TOPICS IN INDEPENDENT COMPONENT ANALYSIS, LIKELIHOOD COMPONENT ANALYSIS, AND SPATIOTEMPORAL MIXED MODELING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Benjamin Brewster Risk

August 2015

TOPICS IN INDEPENDENT COMPONENT ANALYSIS, LIKELIHOOD COMPONENT ANALYSIS, AND SPATIOTEMPORAL MIXED MODELING

Benjamin Brewster Risk, Ph.D.

Cornell University 2015

This dissertation explores dependence patterns using a range of statistical methods: from estimating latent factors in multivariate analysis to mixed modeling of spatially and temporally dependent data. The methods may be applied to many scientific problems and types of data, but here we focus on the application to functional magnetic resonance imaging (fMRI).

In the first chapter, we examine differences between independent component analyses (ICAs) arising from different assumptions, measures of dependence, and starting points of the algorithms. ICA is a popular method with diverse applications including artifact removal in electrophysiology data, feature extraction in microarray data, and identifying brain networks in functional magnetic resonance imaging (fMRI). ICA can be viewed as a generalization of principal component analysis (PCA) that takes into account higher-order cross-correlations. Whereas the PCA solution is unique, there are many ICA methods–whose solutions may differ. Infomax, FastICA, and JADE are commonly applied to fMRI studies, with FastICA being arguably the most popular. A previous study demonstrated that ProDenICA outperformed FastICA in simulations with two components. We introduce the application of ProDenICA to simulations with more components and to fMRI data. ProDenICA was more accurate in simulations, and we identified differences between biologically meaningful ICs from ProDenICA versus other methods in the fMRI analysis. ICA methods require non-convex optimization, yet current practices do not recognize the importance of, nor adequately address sensitivity

to, initial values. We found that local optima led to dramatically different estimates in both simulations and group ICA of fMRI, and we provide evidence that the global optimum from ProDenICA is the best estimate. We applied a modification of the Hungarian (Kuhn-Munkres) algorithm to match ICs from multiple estimates, thereby gaining novel insights into how brain networks vary in their sensitivity to initial values and ICA method. The manuscript resulting from this research is co-authored by David Matteson, David Ruppert, Ani Eloyan (Johns Hopkins University), and Brian Caffo (Johns Hopkins University).

In the second chapter, we develop a new approach for dimension reduction and latent variable estimation by maximizing a non-Gaussian likelihood. Independent component analysis (ICA) is popular in many applications, including cognitive neuroscience and signal processing. Due to computational constraints, principal component analysis is used for dimension reduction prior to ICA (PCA-ICA), which could remove important information. To address this issue, we propose likelihood component analysis (LCA) in which dimension reduction and latent variable estimation is achieved simultaneously by maximizing a likelihood with Gaussian and non-Gaussian components. We present a parametric model using the logistic density and a semi-parametric version using tilted Gaussians with cubic B-splines. We implement an algorithm scalable to datasets common in applications (e.g., hundreds of thousands of observations across hundreds of variables with dozens of latent components). In simulations, our methods recover latent components that are discarded by PCA-ICA methods. PCA-ICA is a popular technique to identify artifacts in functional magnetic resonance imaging. We apply our method to an experiment from the Human Connectome Project with state-of-the-art temporal and spatial resolution, and identify an artifact using LCA that was missed by PCA-ICA. Our results suggest that likelihood component analysis can detect novel signals in neuroimagery.

The third chapter is a departure from the previous topics as it develops a model with Gaussian assumptions. Function magnetic resonance imaging (fMRI) can be used to locate which areas of the brain are activated from thoughts and/or behaviors. In order to assess activation, fMRI data are analyzed by fitting univariate models at every location in the brain, which is called the massive univariate approach. Prior to fitting these models, fMRI data are smoothed for two reasons: to increase the power to detect activated locations and to increase the overlap of corresponding features. However, this decreases the precision with which activation is localized. There is no clear answer to how much smoothing should be used. Moreover, technological improvements that increase the resolution of fMRI data can not be used to increase the resolution of localization if too much smoothing is used. We propose a spatiotemporal mixed model that chooses smoothing in a principled manner that balances its costs and benefits. The model includes a vertex random effect common to all subjects that captures local deviations from regional activation, which obviates the need for smoothing to increase power. The model also includes a subject-vertex random effect that allows subject-specific deviations from the population-level activation, which obviates the need for smoothing to increase the overlap between features in different subjects. We apply our method to high resolution ($2 \times 2 \times 2$ mm) and high frequency (0.72 seconds between scans) fMRI data from the Human Connectome Project and demonstrate the ability to automate smoothing via a unified spatiotemporal mixed model involving a covariance matrix with dimensions 326 million by 326 million.

BIOGRAPHICAL SKETCH

Benjamin Risk wanted to be an ornithologist when he was a child, and this aspiration ultimately led him to the field of statistics. At Dartmouth College, he majored in Environmental and Evolutionary Biology. He completed an honors thesis on the age and habitat-specific demography of the Black-Throated Blue Warbler. The analysis required models for fitting count data containing many zeros with repeated measurements on individuals present in multiple years. With little statistical training, he felt he was unable to adequately address the questions presented by the data. After completing his undergraduate degree, Risk was an analyst at Charles River Associates in Oakland, CA, where he worked on energy economics, environmental economics, and antitrust litigation. Risk's primary duties included coding statistical models in STATA and SAS for expert witnesses. Risk returned to ecology at the University of California, Berkeley, where he focused on conservation biology and avian ecology under the mentorship of Steve Beissinger. During a course on biological modeling geared towards biologists with little mathematical background, Risk realized more mathematical training would be vital to developing quantitative approaches. With the support of his advisor and an applied mathematician, Perry de Valpine, Risk developed a Bayesian model of extinction and colonization dynamics of Black Rail populations in the Sierra Foothills. He found his love for research increasingly revolving around the statistical methods used to conduct scientific inference. Risk then changed career paths and pursued graduate studies in statistics. His previous career included fieldwork on Pinta Island in the Galápagos, the outer foothills of the Sierra Nevada, the cloud forests of the eastern slope of the Andes, the White Mountains of New Hampshire, the Rideau Lakes region in Southern Ontario, and the forest preserves near his hometown in Northbrook, IL. His research currently focuses on statistical methods for analyzing neuroimaging. Risk's hobbies include cycling and guitar. During his time at Berkeley, he told a colleague, "I think I could be a

really good biologist, but I also think I could be an okay statistician." This dissertation is the realization of the latter part of that statement.

ACKNOWLEDGEMENTS

Thank you to my family: my mother Margo, my father Jay, my brothers Jon and Ted, my sister Julia, my stepmother Paula, and my niece Ashlyn. Thank you to my advisors, David Ruppert and David Matteson, for their patience and all they have taught me; thank you to my committee members, Jim Booth and Jacob Bien; thank you to my Cornell statistics colleagues including Didier Chételat, Irina Gaynanova, Will Nicholson, Lucas Mentch, Jón Steingrimsson, Kerstin Frailey, Luo Xiao, Maximillian Chen, Yue Zhao, David Sinclair, and Dan Kowal; thank you to Sara Kaiser; thank you to Nathan Spreng; thank you to Brian Caffo and Ani Eloyan; thank you to Jamie Sorrentino and Anna Matusiewicz; thank you to Steve Beissinger and Perry de Valpine; thank you to Elizabeth Hunter and Orien Richmond; thank you to Jerry Tecklin; thank you to Richard Holmes, Matt Ayres, and Jenn Barg; thank you to John Motsinger and Miles Harrigan.

Chapter 1 was supported in part by grants R01EB012547 and P41EB015909 from the National Institute of Biomedical Imaging and Bioengineering and grant R01NS060910 from the National Institute of Neurological Disorders and Stroke.

Data in Chapters 2 and 3 were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

AN EVALUATION OF INDEPENDENT COMPONENT ANALYSES WITH AN APPLICATION TO RESTING-STATE FMRI

1.1 Introduction

In independent component analysis (ICA), multivariate observations are linearly transformed to minimize dependencies between variables resulting in so-called independent components (ICs). The goal of ICA is to identify both the mixing matrix and the ICs, and the problem is not identifiable if more than one component has a Gaussian distribution (Comon, 1994). ICA has diverse applications including artifact removal in electrophysiology (Iriarte et al., 2003), extracting gene expression features in microarray data (Kong et al., 2008), facial recognition (Bartlett et al., 2002), and separating mixed audio signals (Bell and Sejnowski, 1995). In addition, it has been used in thousands of studies to identify brain networks from functional magnetic resonance imaging (fMRI) (Beckmann, 2012). In fMRI studies, the blood oxygen level dependent (BOLD) signal is an aggregate measure of neural activity across many brain networks that is measured across time. In spatial ICA, the BOLD signal is decomposed into a mixing matrix containing the temporal loadings of ICs and into ICs representing spatial networks. The spatial networks may capture distinct functionalities (e.g., somato-motor, auditory, or visual network), physiological processes (e.g., breathing, heart-beating), and/or artifacts (e.g., head movement) (Damoiseaux et al., 2006).

Networks and their loadings estimated via an ICA contribute to our understanding of the human brain. Recently, there has been a collaborative effort to make a large amount of resting-state fMRI (rs-fMRI) publicly available (Biswal et al., 2010). The BOLD signals in rs-fMRI are measured in subjects who are assigned no particular task,

which contrasts with experimental (i.e., task-based) fMRI. Group ICA can be used to combine data from hundreds of subjects (Calhoun et al., 2001). Coupled with basic biological assumptions regarding spatial contiguity of networks and association with paradigm-related fMRI, group ICA can greatly facilitate the evaluation of resting-state brain networks. The resulting weight matrices of group ICA are often used in inference, for example to compare diseased and non-diseased populations. ICA has been used to identify abnormalities and biomarkers of disorders including Alzheimer's disease (Celone et al., 2006), major depression (Veer et al., 2010), and schizophrenia (Jafri et al., 2008). ICA will likely play an integral role in the Human Connectome Project, which seeks to create a database of all neurological pathways to further our understanding of disease, brain development, and aging (Beckmann, 2012). Many ICA methods exist, and disentangling the differences between methods could improve the ability to use ICA for clinical application and biomarker development.

ICA is a semiparametric problem with a finite dimensional matrix parameter and infinite dimensional IC distributions. Since the IC distributions are latent, one challenge is to find an estimator that is accurate for a wide variety of source distributions. Parametric ICA methods such as information maximization (Infomax) (Bell and Sejnowski, 1995) and FastICA (Hyvarinen, 1999) assume a parametric source distribution and/or properties of higher order moments to derive comparatively simple algorithms. Infomax assumes a distribution (typically logistic), and FastICA assumes a quasi-likelihood function (typically the negative of the hyperbolic cosine). Both methods are commonly used in fMRI studies in part because they are fast even for large datasets. Although these algorithms work well for a variety of IC distributions, a large mismatch between the assumed densities and the true densities can result in inconsistent estimates of ICs (Cardoso, 1998). A semiparametric approach to modeling ICs, product density ICA via tilted Gaussians (ProDenICA), outperformed FastICA in simulations for a large class

of IC distributions (Hastie and Tibshirani, 2003). Other methods using nonparametric estimation of the IC densities have also been developed (Eloyan and Ghosh, 2013; Chen and Bickel, 2006). These are similar to ProDenICA but typically computationally more expensive.

From a biological perspective, a voxel (volumetric pixel) might be a non- or primary contributor to a network, suggesting the use of mixture distributions for some networks in spatial ICA of fMRI (Guo, 2011). Two to three mixtures of normals for an IC has been found to work well in task-based fMRI, where voxels can be regarded as activated by the experiment or their fluctuations may correspond to background noise (Beckmann and Smith, 2004). Simulation studies with two ICs found that FastICA performed poorly when densities were a mixture of normals, while semiparametric methods performed well (Hastie and Tibshirani, 2003; Eloyan and Ghosh, 2013). However, the performance of FastICA, Infomax, and ProDenICA has not been evaluated in dimensions typically found in fMRI applications (e.g., twenty components). Moreover, ProDenICA has not been applied to ICA of fMRI. This suggests a need to determine whether ProDenICA outperforms FastICA and Infomax in simulations with higher dimensions and whether ProDenICA differs from other methods when applied to fMRI.

ICA methods require non-convex optimization, yet fMRI toolboxes that address the issue of sensitivity to initial values are problematic, and most statistical packages do not address the issue whatsoever. A method called Icasso uses agglomerative hierarchical clustering on absolute correlations to match ICs from multiple starting values (Himberg et al., 2004), and the centroids of tight clusters from multiple initializations are regarded as the best estimates for reliable ICs from fMRI (Correa et al., 2007). ICs that do not tightly cluster spatially are excluded from subsequent analyses. Consequently, this approach may mistakenly exclude ICs from analyses of neurological disorders simply

because they have local optima.

The global maximum usually corresponds to the best estimate of the true mixing matrix when an ICA method is statistically consistent (e.g., Matteson and Tsay 2013); unfortunately, some ICA methods are not consistent for many IC distributions. For the FastICA estimator, the set of local maxima of the expected value of the objective function contains the true unmixing matrix under certain conditions relating the true and hypothesized densities, which is referred to as *local consistency* (Hyvarinen, 1999). However, the FastICA objective function does not provide a way to identify which local maximum corresponds to a consistent estimator (if one exists). To investigate, we simulate distributions with large samples sizes and present examples where a local maximum corresponds to the true unmixing matrix but the global maximum is spurious. We also conduct simulations with five, ten, and twenty components and show none of the local optima of FastICA or Infomax that we located are close to the true unmixing matrix, thereby identifying reasonable distributions for which these estimators perform poorly.

It is well known that ICs are only identifiable up to scaled permutations, which is sometimes called the *permutation problem*. As a result, ICs from different initializations or methods are difficult to compare. In contrast to fMRI studies relying upon Icasso (e.g., Correa et al. 2007) or upon matching by highest absolute correlation (e.g., Guo 2011), we optimally match components from different methods via a modification of the Hungarian (Kuhn-Munkres) algorithm (Tichavsky and Koldovsky, 2004). This allows a more detailed comparison of ICs within each method that vary due to initialization, as well as a comparison of ICs between methods that vary due to their assumptions and dependency measures.

To quantify the practical impacts of initialization and choice of methodology, we consider a large collection of rs-fMRI data from multiple data collection centers world-

wide on children and adolescents with Attention Deficit Hyperactive Disorder (ADHD) (Milham et al., 2012). This data was made publicly available in a competition on automated diagnosis of ADHD, and two of the authors of this paper were part of the declared winning team (Eloyan et al., 2012). Here, we use the dataset as a source of multi-subject, multi-site rs-fMRI. We use the group ICA of Calhoun et al. (2001), which is easily adapted to any ICA algorithm and to multi-site rs-fMRI. We evaluate the impact of initial values and compare the mixing matrices and group ICs estimated using FastICA, Infomax, joint approximate diagonalization of eigenmatrices (JADE; Cardoso and Souloumiac 1993), and ProDenICA.

In Section 2, we describe the noise-free ICA model and characterize the objective functions used by FastICA, Infomax, JADE, and ProDenICA. We formalize group ICA as a noisy ICA model with a known number of components, and then we discuss a canonical ordering of ICs and the matching algorithm. In Section 3, we demonstrate the existence of spurious global optima in the FastICA and Infomax objective functions– but not ProDenICA–for simulations with large sample sizes and two components. We also show that the FastICA, Infomax, and ProDenICA algorithms are sensitive to initial values for five, ten, and twenty components, and that ProDenICA is the most accurate. In Section 4, we conduct a group ICA of the ADHD-200 sample using the four methods. In Section 5, we conclude that multiple starting values are necessary and that ProDenICA may be more reliable in fMRI studies.

1.2 ICA methods

1.2.1 The noise-free ICA model

Let $\mathbf{Z}_v$ be a random vector in $\mathbb{R}^Q$ with finite second moments. Without loss of generality, assume $\mathrm{E}\,\mathbf{Z}_v = 0$ and $\mathrm{E}\,\mathbf{Z}_v\mathbf{Z}_v' = \mathbf{I}$, where $\mathbf{Z}_v'$ is the transpose of $\mathbf{Z}_v$. Let the mixing matrix, $\mathbf{A}$, be a $Q \times Q$ matrix of full rank, and denote the unmixing matrix as $\mathbf{W}$, which is equal to $\mathbf{A}^{-1}$. Let $\mathbf{S}_v \in \mathbb{R}^Q$ be a random vector in which the components are mutually independent with $\mathrm{E}\,\mathbf{S}_v = 0$ and $\mathrm{E}\,\mathbf{S}_v\mathbf{S}_v' = \mathbf{I}$. The noise-free ICA model is

$$\mathbf{Z}_v = \mathbf{A}\mathbf{S}_v. \tag{1.1}$$

We observe V identically distributed samples of $\mathbf{Z}_v$. Then the goal is to estimate $\mathbf{W}$, which we can then use to estimate the ICs. We briefly describe four methods to estimate $\mathbf{W}$ below.

1.2.2 Mutual information, maximum likelihood, and Infomax ICA

Minimization of mutual information (MI) provides a unifying framework for a variety of ICA methods, including maximum likelihood (ML), Infomax, and negentropy (Cardoso, 1997, 1998). MI measures the Kullback-Leibler divergence between a joint density (assumed to be known) and the product of its marginal densities. Let $F_{\mathbf{S}}$ denote the joint distribution of a random vector $\mathbf{S} \in \mathbb{R}^Q$, and suppose $F_{\mathbf{S}}$ is absolutely continuous with density $f_{\mathbf{S}}(s)$. Let F_{S_q} denote the marginal distribution of the qth component of $\mathbf{S}$ and $f_{S_q}(s)$ the corresponding density. Let $\Theta = \{\boldsymbol{s} \in \mathbb{R}^Q : f_{\mathbf{S}}(\boldsymbol{s}) > 0\}$. MI is defined as

$$\mathcal{K}(F_{\mathbf{S}}; \prod_{q=1}^{Q} F_{S_q}) = \int_{\boldsymbol{s}\in\Theta} \log\left(\frac{f_{\mathbf{S}}(\boldsymbol{s})}{\prod_{q=1}^{Q} f_{S_q}(s_q)}\right) f_{\mathbf{S}}(\boldsymbol{s})d\boldsymbol{s}. \tag{1.2}$$

Then, $S_1, \ldots, S_Q$ are mutually independent if and only if their MI is equal to zero.

Suppose we have the noise-free ICA model in (1.1) with $\mathbf{W}$ denoting the true unmixing matrix and $F_{\mathbf{S}} = \prod_{q=1}^{Q} F_{S_q}$. Let $\mathcal{O}$ be the set of $Q \times Q$ orthogonal matrices, and let $\mathcal{P}$ be the set of $Q \times Q$ signed permutation matrices. Define the equivalence relation $\mathbf{A} \cong \mathbf{B}$ if there exists some $\mathbf{P} \in \mathcal{P}$ such that $\mathbf{A} = \mathbf{PB}$. Then,

$$\mathbf{W} \cong \underset{\mathbf{O} \in \mathcal{O}}{\operatorname{argmin}} \, \mathcal{K}\left(F_{\mathbf{OZ}}; \prod_{q=1}^{Q} F_{\mathbf{o}_q'\mathbf{Z}} \right),$$

where $\mathbf{o}_q'$ is the qth row of $\mathbf{O}$. Let $\mathcal{H}(\mathbf{S})$ denote the differential entropy,

$$\mathcal{H}(\mathbf{S}) = -\int_{\boldsymbol{s} \in \mathbb{R}^Q} \{\log f(\boldsymbol{s})\} f(\boldsymbol{s}) d\boldsymbol{s}, \tag{1.3}$$

and note that the MI is equal to the sum of the marginal entropies less the joint entropy, $\mathcal{K}(F_{\mathbf{S}}; \prod_{q=1}^{Q} F_{S_q}) = \sum_{q=1}^{Q} \mathcal{H}(S_q) - \mathcal{H}(\mathbf{S})$.

If the true joint density of the ICs is known, we can define the objective function for identically distributed observations $\mathbf{z}_1, \ldots, \mathbf{z}_V$ as

$$\mathcal{J}_{MI}(\mathbf{O}) = -\sum_{v=1}^{V} \sum_{q=1}^{Q} \log f_{S_q}(\mathbf{o}_q'\mathbf{z}_v) + \sum_{v=1}^{V} \log f_{\mathbf{S}}(\mathbf{O}\mathbf{z}_v).$$

Since $\sum_{v=1}^{V} \log f_{\mathbf{S}}(\mathbf{O}\mathbf{z}_v)$ is invariant to rotations $\mathbf{O}$, we obtain

$$\widehat{\mathbf{W}} = \underset{\mathbf{O} \in \mathcal{O}}{\operatorname{argmin}} \; -\sum_{v=1}^{V} \sum_{q=1}^{Q} \log f_{S_q}(\mathbf{o}_q'\mathbf{z}_v). \tag{1.4}$$

From (1.4), it is clear that the MI criterion is equal to the negative of the ML criterion.

In practice, the densities f_{s_q} are not known, so most ML ICA methods assume a parametric density $f_{s_q}^*$. In particular, the Infomax criterion is equal to the ML criterion in which the information transfer function described in Bell and Sejnowski (1995) equals the (assumed) common cumulative distribution function of the ICs (Cardoso, 1997). The information transfer function is most commonly taken to be the logistic distribution,

$F^*_{s_q}(x) = 1/(1+e^{-x})$ for $x \in \mathbb{R}$, and $\widehat{\mathbf{W}}$ is not restricted to $\mathcal{O}$ (Bell and Sejnowski, 1995). Let $\mathcal{B}$ be the set of full rank $Q \times Q$ matrices, and let $\mathbf{B} \in \mathcal{B}$ with rows $\mathbf{b}'_q$. Then the infomax objective function is

$$\mathcal{J}_{Info}(\mathbf{B}) = V \log|\det \mathbf{B}| + \sum_{v=1}^{V}\sum_{q=1}^{Q}\{\log f^*_s(\mathbf{b}'_q\mathbf{z}_v)\}.$$

The wrong $F^*_{\mathbf{S}}$ may still result in a consistent estimator of $\mathbf{W}$ and successfully recover ICs from a variety of distributions, although the use of $F^*_{\mathbf{S}} \neq F_{\mathbf{S}}$ always results in some loss of efficiency. Cardoso (1998) provided a heuristic treatment of the consistency of ICA estimators, where $\widehat{\mathbf{W}}$ may be inconsistent when there is a large mismatch between the hypothesized and true IC distributions. Here, we investigate the accuracy of estimators via simulations with large sample sizes, which is suggestive of consistency properties; a formal consistency analysis is beyond the scope of this paper. We modify the Infomax algorithm from Bell and Sejnowski (1995) as described in A.1.1. Our R code is available on request.

1.2.3 Negentropy and the FastICA algorithm

The FastICA algorithm is based on maximizing the sum of the marginal negentropies. Under the constraint of orthogonal ICs, maximizing negentropy is equal to minimizing MI (Hyvarinen, 1999). Using the notation from (1.3), negentropy is defined as

$$\mathcal{I}(\mathbf{X}) = \mathcal{H}(\mathbf{Y}) - \mathcal{H}(\mathbf{X}),$$

where $\mathbf{Y} \sim \mathcal{N}(0, \mathrm{E}\,\mathbf{X}\mathbf{X}^T)$. Note that for $\mathbf{X} \sim (0, \mathbf{I})$, $\mathcal{H}(\mathbf{Y}) = Q\mathcal{H}(Y)$ with $Y \sim \mathcal{N}(0,1)$. Then the MI for linear transformations in (1.2) equals

$$\mathcal{K}(F_{\mathbf{S}}; \prod_{q=1}^{Q} F_{S_q}) = \mathcal{I}(\mathbf{O}\mathbf{Z}) - \sum_{q=1}^{Q}\mathcal{I}(\mathbf{o}'_q\mathbf{Z}).$$

Since multivariate negentropy is invariant to orthogonal rotations, it follows that minimizing MI is equal to maximizing the negentropy of the marginals.

Approximations to marginal negentropy can take the form (Hyvarinen, 1999)

$$\mathcal{I}(X) \propto \left[\mathrm{E}\left\{G\left(X\right)\right\} - \mathrm{E}\left\{G\left(Y\right)\right\}\right]^2, \tag{1.5}$$

where G is a non-quadratic function referred to as the "non-linear function." A common choice is $G(x) = \frac{1}{\alpha}\log\{\cosh(\alpha x)\}$ for $1 \leq \alpha \leq 2$. Then for observations $v = 1, \ldots, V$, define the objective function

$$\mathcal{J}_{FastICA}\left(\mathbf{O}\right) = \sum_{q=1}^{Q}\left[\frac{1}{V}\sum_{v=1}^{V} G\left(\mathbf{o}_q'\mathbf{z}_v\right) - \mathrm{E}\left\{G\left(Y\right)\right\}\right]^2, \tag{1.6}$$

where $\mathrm{E}\{G(Y)\}$ is a known constant. This is maximized using an approximative Newton algorithm, or *fixed-point algorithm* (Hyvarinen, 1999). The fixed-point algorithm assumes a diagonal Hessian matrix, which allows for faster rates of convergence than the Infomax algorithm and fewer computations than an exact Newton algorithm. It can also be derived as a stochastic gradient ascent algorithm for quasi-MLE, where the derivative of G equals the score function (Hyvärinen and Oja, 2000). We implement FastICA using the R package of that name by Marchini et al. (2010) with the log cosh nonlinearity, $\alpha = 1$, and the symmetric estimation scheme.

1.2.4 ProDenICA

ProDenICA combines semiparametric estimation of the IC distributions with a fixed-point algorithm (Hastie and Tibshirani, 2003). The joint density of independent ICs is modeled as the product of tilted Gaussians, $f_{\mathbf{S}}(\boldsymbol{s}) = \prod_{q=1}^{Q}\phi(s_q)e^{g_q(s_q)}$. Here, ϕ is a standard normal density and $g_q(s_q)$ is estimated with cubic B-splines. Let $h_q(x)$ denote

the second derivative of $g_q(x)$. The objective function is a penalized log likelihood,

$$\mathcal{J}_{ProDen}(\mathbf{O}) = \sum_{q=1}^{Q} \frac{1}{V} \left(\sum_{v=1}^{V} \log \phi(\mathbf{o}_q^T \mathbf{z}_v) + g_q(\mathbf{o}_q^T \mathbf{z}_v) \right) - \int \phi(x) e^{g_q(x)} dx - \lambda \int \{h_q(x)\}^2 dx, \tag{1.7}$$

where the first penalty enforces the constraint that $\phi(x)e^{g_q(x)}$ integrates to one, and the second is a roughness penalty.

This objective function is maximized by alternately estimating g_q, which is found using an application of generalized additive models, and updating $\mathbf{O}$ with one-step of the fixed-point algorithm used in FastICA. Since it is the log likelihood ratio of the tilted Gaussian to Gaussian, g_q is used as an estimate of marginal negentropy in the fixed-point algorithm. Thus, ProDenICA adapts to the IC distributions while minimizing dependencies. We implement ProDenICA using the R package of that name by Hastie and Tibshirani (2010), and we describe solutions to computational issues that arose when using ProDenICA in A.1.2.

1.2.5 JADE

For mutually independent random variables, the cross cumulants of all orders are equal to zero. JADE seeks a rotation of whitened data that approximately diagonalizes the fourth-order cross-cumulant tensor (Cardoso and Souloumiac, 1993). The JADE algorithm requires all but one of the excess kurtoses to be non-zero, and it is based on necessary, but not sufficient, conditions for independence. An important difference between JADE and other algorithms is that it does not require initialization. We implement JADE using the R package of that name by Nordhausen et al. (2011).

1.2.6 A group ICA model

We estimate group ICs using the approach proposed by Calhoun et al. (2001), which involves a two-stage dimension reduction via the singular value decomposition prior to applying a noise-free ICA. Let s_{vq}, $q \in 1, \ldots, Q$, denote mutually independent random variables and $\boldsymbol{s}_v = [s_{v1}, \ldots, s_{vQ}]'$. We assume $\boldsymbol{s}_v$ are iid F for $F \in \mathcal{F}$, in which $\mathcal{F}$ is the class of Q-variate non-Gaussian mean zero distributions with covariance equal to the identity matrix. Let $\mathbf{M}^{(m)}$ be a $T_r \times Q$ matrix of mixing weights for the ICs for the mth subject. Our probabilistic spatial group ICA model is

$$\mathbf{x}_v^{(m)} = \mathbf{M}^{(m)} \boldsymbol{s}_v + \boldsymbol{\epsilon}_v^{(m)}, \tag{1.8}$$

where $\boldsymbol{\epsilon}_v^{(m)}$ has mean zero and is the error that is not explained by the group ICs.

Suppose $\mathbf{X}^{(m)}$ is a $V \times T_r$ matrix where each column corresponds to a three-dimensional snapshot of the BOLD signal that has been vectorized, and suppose the data have been centered such that both rows and columns have zero mean. Now, consider the singular value decomposition (SVD) of observations from subject m: $\mathbf{X}^{(m)} = \widehat{\mathbf{U}}^{(m)}\widehat{\mathbf{D}}^{(m)}\widehat{\mathbf{V}}^{(m)}$. Let $\widehat{\mathbf{U}}_Q^{(m)}$ denote the first Q left singular vectors and $\widehat{\mathbf{Z}}_Q^{(m)} = \sqrt{V}\widehat{\mathbf{U}}_Q^{(m)}$, where $\sqrt{V}$ standardizes $\widehat{\mathbf{Z}}_Q^{(m)}$ to have sample covariance equal to the identity matrix.

We can align the voxels across subjects from multiple sites, while in general we cannot align time courses in rs-fMRI. Consequently, we concatenate the data matrices $\widehat{\mathbf{Z}}_Q^{(m)}$ across subjects into a matrix $\mathbf{Y}$ with dimensions $V \times MQ$. Next, a second SVD is performed, and the first Q^* left singular vectors are retained and multiplied by $\sqrt{V}$. Here, we let $Q^* = Q$. This results in a whitened data matrix $\widehat{\mathbf{Z}}$ with dimensions $V \times Q$. Applying the methods described in Section 2, we now find a linear transformation $\widehat{\mathbf{W}}$ that results in group ICs $\widehat{\mathbf{S}}$ that minimize a measure of dependence. Thus, the multi-

subject ICA problem is reduced to the noise-free ICA model in (1.1).

Note that we can estimate $\mathbf{M}^{(m)}$ for each subject using standard multivariate regression, such that for a given $\mathbf{S}$, we use least squares to solve $\mathbf{X}^{(m)} = \mathbf{S}\mathbf{M}^{(m)'} + \mathbf{E}^{(m)}$, where $\mathbf{E}^{(m)}$ is the $V \times T_r$ matrix of residuals not accounted for by the group components. With this approach, any ICA method can be applied to fMRI from multiple subjects and sites.

1.2.7 Canonical form for ICA and matching ICs

The ICA model as presented in (1.1) is only identifiable on an equivalence class of signed permutations since both $\mathbf{W}$ and $\mathbf{S}$ are unknown (Section 1.2.2). Eloyan and Ghosh (2013) demonstrate that the ICA model is uniquely identified if $\mathrm{E}\, s_1^3 > \cdots > \mathrm{E}\, s_Q^3 \geq 0$. Since $\mathrm{E}\, s_q = 0$ and $\mathrm{E}\, s_q^2 = 1$, this is the same as assuming the skewnesses are distinct and positive. Then we define a canonical form for the ICs:

Definition 1. *Let $\widehat{\gamma}_q$ denote the sample skewness for the qth IC. Then the canonical form for $\widehat{\mathbf{S}}$ is the signed permutation that results in $\widehat{\gamma}_1 > \cdots > \widehat{\gamma}_Q \geq 0$.*

In fMRI, assuming positive skewness is biologically plausible because voxels that have very positive BOLD signals may be considered primary contributors to a network, and in practice, many IC densities have large skewnesses.

We matched ICs from multiple estimates using a modification of the Hungarian algorithm proposed by Tichavsky and Koldovsky (2004). For two estimates $\widehat{\mathbf{S}}_{(1)}$ and $\widehat{\mathbf{S}}_{(2)}$, let $\hat{s}_i^{(1)}$ and $\hat{s}_j^{(2)}$ be the ith and jth columns, respectively. Let $\|\cdot\|$ denote the L2 norm. Let $\mathbf{C}$ be the cost matrix with elements defined by an auxiliary metric $c_{i,j} = \min(\|\hat{s}_i^{(1)} - \hat{s}_j^{(2)}\|, \|\hat{s}_i^{(1)} + \hat{s}_j^{(2)}\|)$, which accounts for the sign ambiguity. Let $\mathcal{S} = \{\sigma : \sigma = \{\sigma(1), \ldots, \sigma(Q)\}$ be the set of all permutations $\{1, \ldots, Q\}$. The Hungarian

algorithm is used to find the permutation σ^* such that

$$\sigma^* = \underset{\sigma \in \mathcal{S}}{\operatorname{argmin}} \sum_{i=1}^{Q} c_{i,\sigma(i)}.$$

Details are described in A.2. R code implementing the canonical form and the matching algorithm is available by request.

1.3 Simulation study

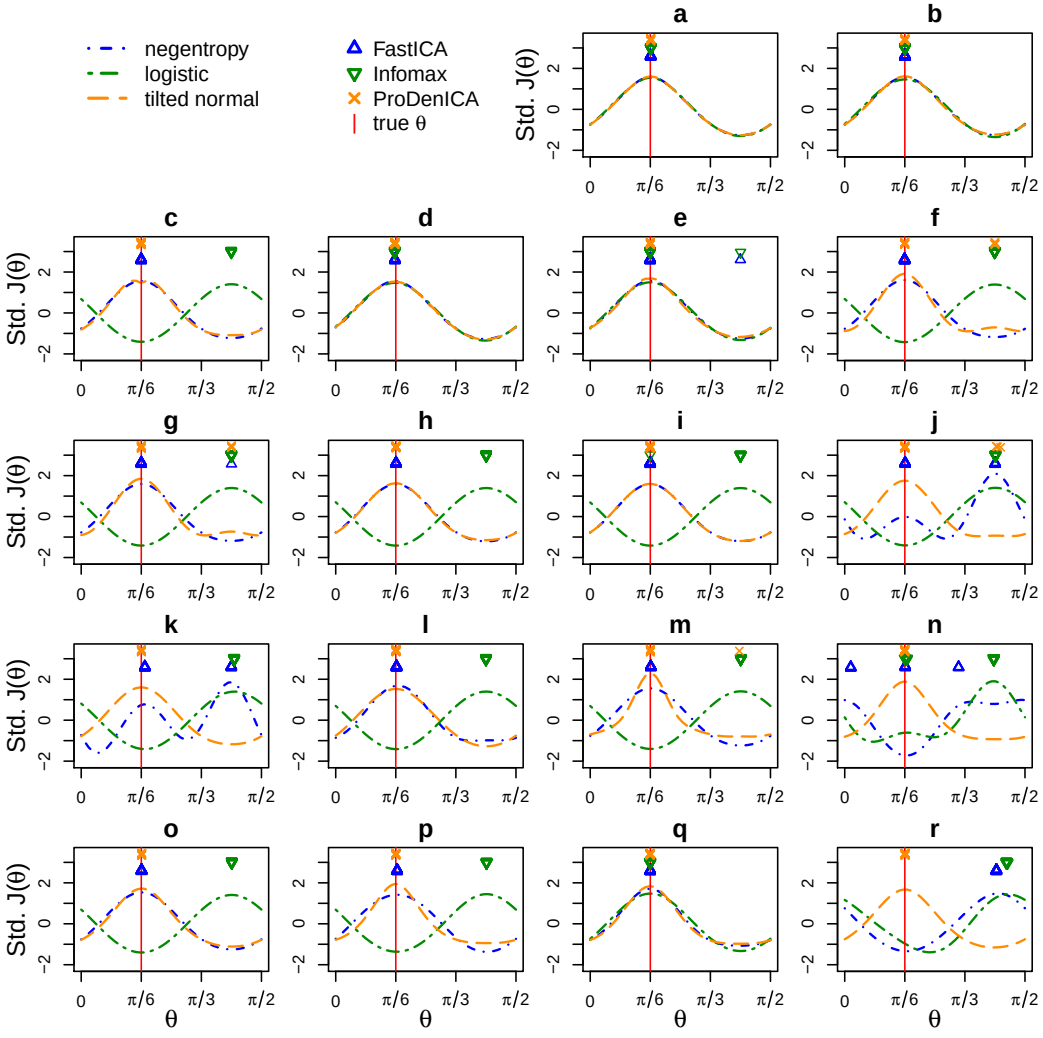
1.3.1 Convexity and accuracy for $Q = 2$

We simulated pairs of identically distributed ICs for eighteen distributions that were used in previous ICA studies (Bach and Jordan, 2003; Hastie and Tibshirani, 2003) including the t-distribution, exponential, double exponential, uniform, a mixture of exponentials, and various symmetric and asymmetric mixtures of normals (A.1.3; Figure A.1).

First, we examined the objective functions for two components for each distribution. We defined $\mathbf{W}$ using the Givens parameterization with $\theta_{true} = \pi/6$. For each distribution, we conducted one simulation with a very large sample size (V=131,072), such that inaccuracies would be suggestive of consistency issues rather than chance variability or small-sample bias. We evaluated the objective functions on a grid for $\theta \in [0, \pi/2]$ with mesh size $\pi/100$. Then for each estimator we estimated $\hat{\theta}_i$ using $N = 25$ equally spaced starting values in $[0, \pi/2]$.

For FastICA and ProDenICA, there are distributions for which the objective functions include local maxima (Figure 1.1). For the symmetric, unimodal, and super-Gaussian (having positive excess kurtosis) distributions a, b, and d (t-distribution with

Figure 1.1: Objective functions (standardized $J(\theta)$; lines) for $V = 131,072$ and $Q = 2$ from distributions *a*-*r* (see Figure A.1) using the angular (Givens) parameterization with $\theta_{true} = \pi/6$ and $\theta \in [0, \pi/2]$ and parameter estimates (characters; y-value chosen for display purposes) from 25 initial values equally spaced in $[0, \pi/2]$.



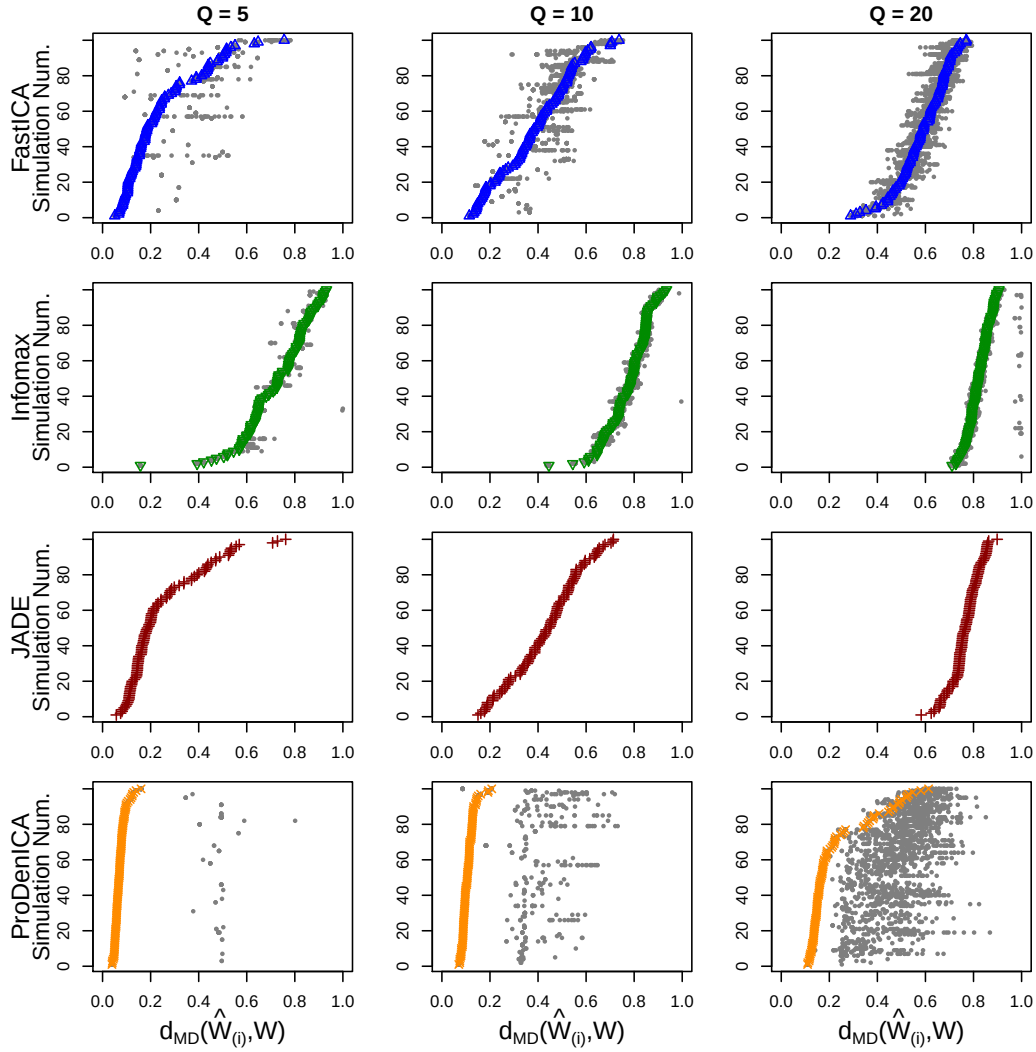
$df = 3$, double exponential, and t-distribution with $df = 5$, respectively), the global maximum for each method correctly identifies θ_{true}, and there are no complications owing to local maxima. In contrast, the asymmetric mixture of two normals in distribution k contains a local maximum for both FastICA and ProDenICA. Thus, even when $Q = 2$, local maxima can be an issue.

It also appears that Infomax and FastICA typically and occasionally, respectively, identify the wrong optima, while the global maxima for ProDenICA correctly identify θ_{true}. The global maxima is associated with θ_{true} for all methods for distributions a, b, d, and e, which are all super-Gaussian, unimodal distributions. For sub-Gaussian distributions f through r, the minimum of Infomax, rather than a maximum, usually appears to correspond to θ_{true} (the one exception is distribution q, which has the largest kurtosis among all sub-Gaussian distributions examined). This is indicative of the Infomax estimator being inaccurate for sub-Gaussian distributions (see Lee et al. 1999). FastICA misidentifies θ_{true} for distributions j and k, which are asymmetric mixtures of normals. For these distributions, a local maximum is associated with θ_{true}, but the global maximum suggests that the FastICA estimator is not consistent. Additionally, $theta_{true}$ in distribution r is associated with a minimum of the FastICA objective function instead of a maximum, which suggests the FastICA method may not be locally consistent for some mixtures of normals.

1.3.2 Convexity and accuracy for $Q = 5$, 10, and 20

To examine convexity and accuracy in higher dimensions, we conducted 100 simulations of the ICA model in (1.1) for $Q = 5$, 10 and 20 randomly chosen (with replacement) distributions from those in Figure A.1. We used 25 initial values generated via latin

Figure 1.2: Simulations using $Q = 5$, 10, or 20 from randomly chosen distributions with $V = 1,024$. For $k =$ FastICA, Infomax, and ProDenICA, the results from 25 initial values for 100 simulations are depicted: small gray points correspond to stationary points ($\widehat{\mathbf{W}}^k_{(i)}, i = 1, \ldots, 25$), and symbols correspond to the global maximum ($\widehat{\mathbf{W}}^k_{(0)}$). For each method k, simulations are sorted from lowest to highest $d_{MD}(\widehat{\mathbf{W}}^k_{(0)}, \mathbf{W})$. The JADE algorithm is not initialized with multiple values.



hypercube sampling of the rotation angles for each simulation, as described in the A.1.3. We used the minimum distance (d_{MD}) measure introduced in Ilmonen et al. (2010) and defined in A.1.4. Let $\widehat{\mathbf{W}}_{(i)}$ denote the unmixing matrix estimated from the ith initial value, $i = 1, \ldots, N$. We then examined $d_{MD}(\widehat{\mathbf{W}}_{(i)}, \mathbf{W})$.

From these simulations, the methods ordered from most to least accurate were ProDenICA, FastICA, JADE, and Infomax (Figure 1.2).The MD measure tended to increase as the number of components increased, although this is partly owing to the manner in which d_{MD} scales with dimension.

Infomax was inaccurate in part because it performs poorly for sub-Gaussian distributions, and fourteen of the eighteen distributions in Figure A.1 are sub-Gaussian. We also investigated the performance of the methods when all ICs had a logistic distribution, which is the best-case scenario for Infomax. Using ten components and the simulation design described above, the means ± standard errors of d_{MD} for FastICA, Infomax, JADE, and ProDenICA were 0.273 ± 0.007, 0.263 ± 0.005, 0.377 ± 0.008, and 0.350 ± 0.014. Not surprisingly, in the unlikely situation where the IC distributions are known, there is a benefit to using the true likelihood in (1.4) rather than the semi-parametric likelihood in (1.7).

For FastICA, two issues are clear from Figure 1.2: there are many stationary points, and in most instances, there exist stationary points that are closer than the global minimum to the true unmixing matrix. Regarding the first issue, comparing the negentropy approximations (1.6) from many initial values would eliminate the use of estimators to the right of the global maximum. The second issue is more problematic. Ideally, we would like to identify the left-most stationary point as our estimate rather than the global maximum. This left-most point represents an *empirical oracle* since it is only known when θ_{true} is known. Given the local consistency properties of FastICA (Hyvarinen, 1999), it is not surprising that there exist local maxima that are closer to the true unmixing matrix than the FastICA solution. But in FastICA, the left-most gray point is often still a poor estimate of $\mathbf{W}$.

In contrast to the other methods, the ProDenICA global maximum usually corre-

sponded to the left-most gray point, and ProDenICA clearly dominated all other estimators (Figure 1.2). ProDenICA is computationally more expensive than other methods (Table A.1). For $Q = 20$ and V =1,024, ProDenICA, Infomax, FastICA, and JADE took approximately 9 minutes, 25 seconds, 7.5 seconds, and 4 seconds, respectively.

1.4 Group ICA of resting-state fMRI

1.4.1 Resting-state fMRI dataset

Data were selected for analysis from the ADHD-200 Data Sample (Milham et al., 2012), which consists of rs-fMRI data from children and adolescents (ages 7-21) from 8 sites comprising 491 typically developing subjects and 285 with ADHD (Table A.2). The number of time slices recorded varied by site from 76 to 261. We restricted our analysis to subjects that were right-hand dominant with no history of drug therapy and to images with no quality control flags. This resulted in 206 typically developing and 78 ADHD children and adolescents from four sites (Table A.3). Data were registered and masked using the MNI 152 T1 3 mm template. Processing scripts were based on the 1,000 Functional Connectome project's (Biswal et al., 2010) processing scripts. We aggregated adjacent voxels to result in $6 \times 6 \times 6$ mm voxels. Additional information is provided in A.3.1.

To determine the number of components, rs-fMRI studies frequently fix the number of ICs at twenty, which is sufficient to capture the most frequently observed large-scale resting-state networks (Smith et al., 2009). Task-based fMRI studies sometimes use a probabilistic PCA (PPCA) prior to ICA to determine the number of ICs (Beckmann and Smith, 2004). The signal-to-noise ratio is smaller in rs-fMRI than task-based fMRI, and

a low signal-to-noise ratio can be problematic for PPCA. Consequently, we followed previous studies and let $Q = Q^* = 20$.

1.4.2 Differences within algorithms

We examined the sensitivity to initialization of FastICA, Infomax, and ProDenICA on group ICA of the ADHD-200 dataset. We generated $N = 1{,}000$ initial values using latin hypercube sampling of the Givens rotation angles. We created a dissimilarity matrix with entries $d_{MD}(\widehat{\mathbf{W}}^k_{(i)}, \widehat{\mathbf{W}}^k_{(j)})$ for the kth method, $i \neq j \in 1, \ldots, 1000$, and $Q = 20$. We also created a dissimilarity matrix for each IC. Define

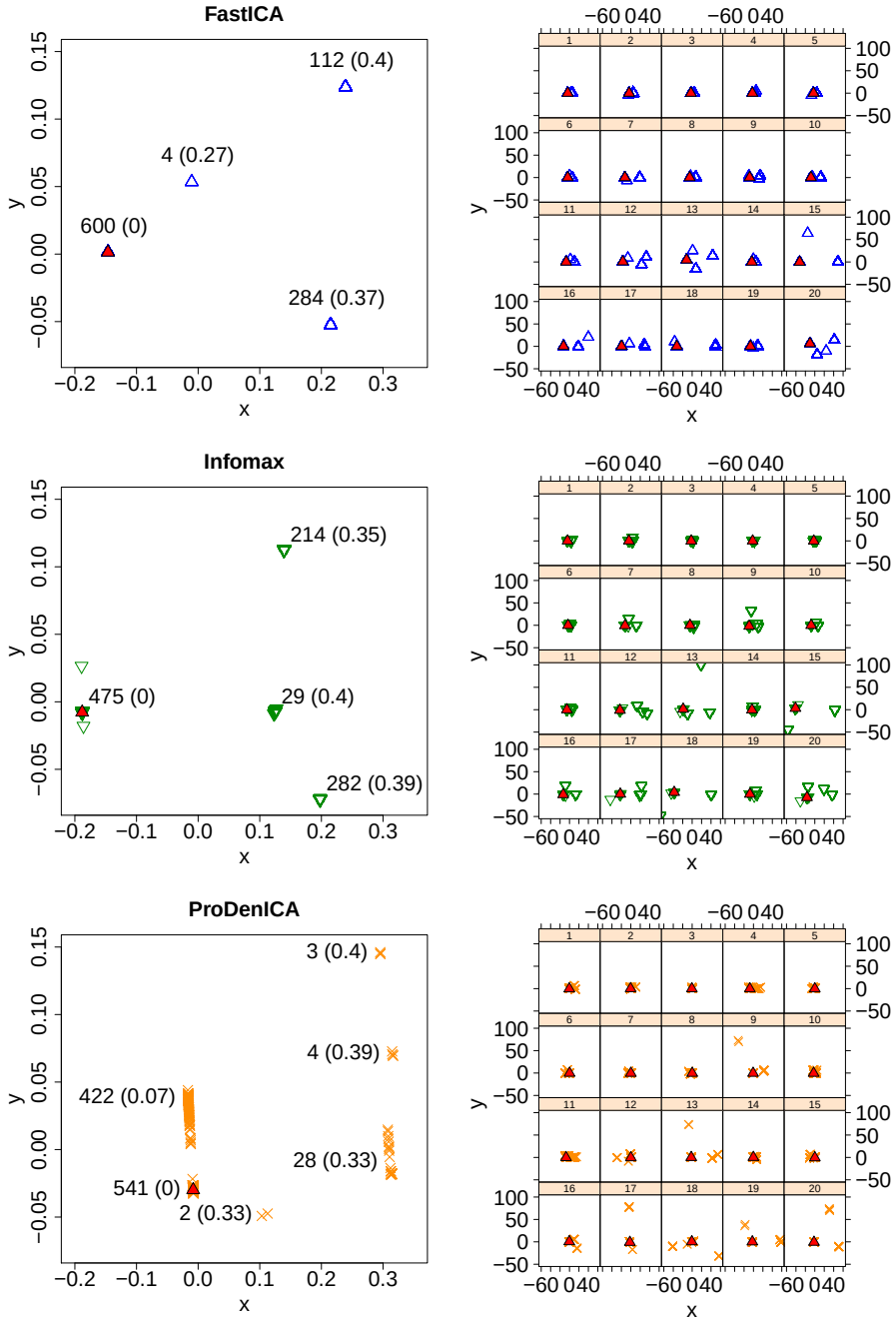
$$\widehat{\mathbf{W}}^k_{(0)} = \underset{i \in 1,\ldots,N}{\operatorname{argmax}} \mathcal{J}(\widehat{\mathbf{W}}^k_{(i)}).$$

Let $\widehat{\mathbf{S}}^k_{(0)}$ be the estimated ICs associated with $\widehat{\mathbf{W}}^k_{(0)}$ and ordered as in Definition 1. Define $\widehat{\mathbf{S}}^k_{(i)}$, $i = 1, \ldots, N$, to be the ICs associated with $\widehat{\mathbf{W}}^k_{(i)}$ that have been matched to $\widehat{\mathbf{S}}^k_{(0)}$. The dissimilarity matrix for the qth IC has entries $\|\widehat{\mathbf{S}}^k_{(i),q} - \widehat{\mathbf{S}}^k_{(j),q}\|_2$, in which $\widehat{\mathbf{S}}^k_{(i),q}$ is the qth column of $\widehat{\mathbf{S}}^k_{(i)}$. We then used classical multidimensional scaling (Torgerson, 1952) with two dimensions to visualize the dissimilarities among unmixing matrices.

In estimates of the mixing matrix, there were four basins of attraction for both FastICA and Infomax (Figure 1.3). For ProDenICA, there were two major basins of attraction and four smaller basins. In all methods, the basin with the most points contained the argmax, along with 60%, 47.5%, and 54.1% of estimates for FastICA, Infomax, and ProDenICA, respectively. The remaining 40% of FastICA estimates had an MD (relative to $\widehat{\mathbf{W}}^k_{(0)}$) of approximately 0.38; 52.5% of Infomax estimates had an MD of approximately 0.37; and 42% of ProDenICA estimates had an MD of 0.07 and another 3.5% had an MD of approximately 0.34.

In estimates of the individual ICs, it is clear that some ICs were nearly identical

Figure 1.3: Multidimensional scaling of $d_{MD}(\widehat{\mathbf{W}}^k_{(i)}, \widehat{\mathbf{W}}^k_{(j)})$ with the number of points in each basin and the average d_{MD} from the basin to $\widehat{\mathbf{W}}^k_{(0)}$ in parentheses, where k indexes method and $i \neq j \in 1, \ldots, 1000$ (left), and $\|\widehat{\mathbf{S}}^k_{(i),q} - \widehat{\mathbf{S}}^k_{(j),q}\|_2$ for $q = 1, \ldots, 20$ (right). The coordinates of $\widehat{\mathbf{W}}^k_{(0)}$ and $\widehat{\mathbf{S}}^k_{(0),q}$, $q = 1, \ldots 20$, are depicted by solid triangles.

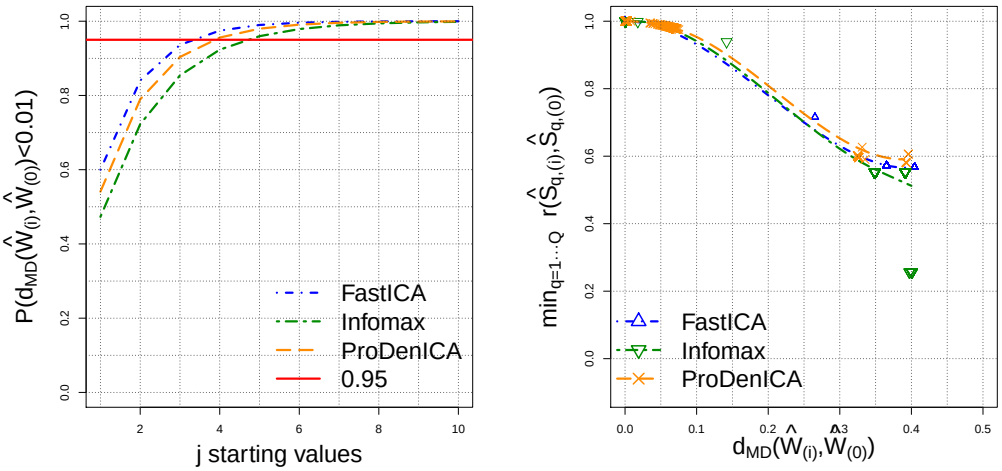


for most starting values (e.g., ICs 1 through 6 for all methods; recall that the ICs are ordered by decreasing skewness; see Figure 1.3), while others were more sensitive to initialization in all methods (e.g., ICs 13, 17, 18, and 20), and some were sensitive in some methods but not others (e.g., IC 15 was sensitive in FastICA and Infomax, but not ProDenICA; IC 19 was sensitive in ProDenICA, but not FastICA or Infomax). Overall, the estimation of ICs with the largest skewnesses and kurtoses (where kurtosis was generally higher in lower-numbered ICs) tended to be more stable than those that were more nearly symmetric with lower kurtoses (see Figure A.2).

We estimated the probability of obtaining $\widehat{\mathbf{W}}^k_{(0)}$ using j starting values. Consider the probability of obtaining an initial value that is close to the argmax, $P(d_{MD}(\widehat{\mathbf{W}}^k_{(0)}, \widehat{\mathbf{W}}^k_{(i)}) < \delta) \geq \epsilon$, when using j starting values. We chose δ such that $\{d_{MD} < \delta\}$ is the event that we have found the global maximum (within some numerical tolerance). Here, we let $\delta = 0.01$. Now recall the hypergeometric distribution, $P(X = x|N, m_k, j) = \{\binom{m_k}{x}\binom{N-m_k}{j-x}\}/\binom{N}{j}$, where N is the total number of starting values (N = 1,000), m_k is the number of times $\widehat{\mathbf{W}}^k_{(i)}$ was within δ of $\widehat{\mathbf{W}}^k_{(0)}$, j is the number of starting values for which we wish to calculate the probability of getting within δ, and x is the number of times that $\widehat{\mathbf{W}}^k_{(i)}$ is within δ of $\widehat{\mathbf{W}}^k_{(0)}$ when using j starting values. We calculated $P(X > 0|N = 1{,}000, m_k, j)$ for $j \in 1, \ldots, 10$. We also calculated $\min_{q=1,\ldots,Q} r(\widehat{\mathbf{S}}^k_{(i),q}, \widehat{\mathbf{S}}^k_{(0),q})$, where r is the Pearson correlation, and examined the relationship of this minimum correlation to d_{MD}.

In our application, FastICA, Infomax, and ProDenICA required 4, 5, and 4 initial values, respectively, to have a greater than 0.95 probability of obtaining the argmax (Figure 1.4). When comparing ICs from different initializations, $\min_{q=1,\ldots,Q} r(\widehat{\mathbf{S}}_{(i),q}, \widehat{\mathbf{S}}_{(0),q})$ was on average approximately 0.60 and as low as 0.25 for $d_{MD} > 0.3$. Overall, a minimum distance measure less than 0.10 for some pair of $\widehat{\mathbf{W}}^k_{(i)}$ and $\widehat{\mathbf{W}}^k_{(0)}$ translated to a minimum correlation (over q) between $\widehat{\mathbf{S}}_{(i),q}$ and $\widehat{\mathbf{S}}_{(0),q}$ of at least 0.95.

Figure 1.4: The probability of obtaining $\widehat{\mathbf{W}}^k_{(0)}$ when using j initial values for k = FastICA, Infomax, or ProDenICA (left). The relationship between $d_{MD}(\widehat{\mathbf{W}}^k_{(i)}, \widehat{\mathbf{W}}^k_{(0)})$ and the Pearson correlation between $\widehat{\mathbf{S}}^k_{(i),q}$ and $\widehat{\mathbf{S}}^k_{(0),q}$ (lines are from a loess smoother), where for each initial value, the symbol denotes the minimum correlation $r(\widehat{\mathbf{S}}^k_{(i),q}, \widehat{\mathbf{S}}^k_{(0),q})$ with $q = 1, \ldots, 20$ versus $d_{MD}(\widehat{\mathbf{W}}^k_{(i)}, \widehat{\mathbf{W}}^k_{(0)})$ (right).



1.4.3 Differences between algorithms

We matched $\widehat{\mathbf{W}}^k_{(0)}$, for k indexing Infomax, JADE, and ProDenICA, to the canonically ordered results from FastICA. We also compared each method to the SVD, which represents a baseline for understanding the impact of the additional rotation via ICA. We compared unmixing matrices using three measures: (1) the MD measure, d_{MD}; (2) the Amari measure (Amari et al., 1996); and (3) the Frobenius norm between matched unmixing matrices.

FastICA and Infomax had very similar results, while ProDenICA and JADE differed from each other and from FastICA and Infomax (A.4). All ICA solutions were substantially different from the SVD solution. The measures between ICA unmixing matrices were all substantially smaller than between random matrices (see A.3.2).

We compared estimated ICs between methods using Pearson correlations, where all methods were matched to the canonically ordered FastICA. We used Kolmogorov-Smirnov (KS) two-sample tests to examine differences in the CDFs of matched ICs. We did not formally test for equality in distribution because IC samples (i.e., values at different voxels) were spatially dependent. Nonetheless, we calculated FDR-adjusted p-values (Benjamini and Hochberg, 1995) as a measure of the difference between ICs, as described in A.3.2. Lastly, we estimated the density of ICs for each method using Gaussian kernels (A.3.2).

The Pearson correlations were high for most ICs but not all (Table 1.1), and the shapes of the estimated densities across the four methods were similar for most distributions with some notable exceptions (Figure A.2). In contrast, the KS statistics often indicated differences in the distributions of ICs by method (Table A.4). Overall, $r(\widehat{\mathbf{S}}^k_{(0),q}, \widehat{\mathbf{S}}^l_{(0),q}) > 0.95$ in 78/120 comparisons (excluding SVD) and $r(\widehat{\mathbf{S}}^k_{(0),q}, \widehat{\mathbf{S}}^l_{(0),q}) < 0.80$ in 12/120 comparisons. Some ICs were highly correlated for all methods (e.g., ICs 1-3, 5-10, 14), while for other ICs, ProDenICA and JADE had relatively low correlations with FastICA and Infomax (e.g., ICs 13 and 20), and occasionally, ProDenICA differed from all other methods (e.g., IC 11) or JADE differed from all other methods (e.g., IC 19). In the KS tests, FDR-adjusted $p \leq 0.01$ in 72/120 comparisons. In some cases, $p \leq 0.01$ even though the ICs were highly correlated (e.g., IC 3). For FastICA and Infomax, $p > 0.05$ for all ICs except IC 4. In cases with low correlations, differences in the density plots were often visible (e.g., in IC 13, ProDenICA was less peaked; also see ICs 3, 4, 18, and 19). Sometimes correlations were high, but KS-statistics and density plots indicated differences between ProDenICA and other methods (e.g., IC 12), or differences between JADE, ProDenICA, and FastICA/Infomax (e.g., IC 3).

A visual comparison of the spatial configuration of the group ICs revealed that mod-

Table 1.1: Pearson correlation between matching ICs for each method from the rs-fMRI study.

Method1	Method2	IC 1	IC 2	IC 3	IC 4	IC 5	IC 6	IC 7	IC 8	IC 9	IC 10
SVD	FastICA	0.51	0.47	0.33	0.35	0.43	0.44	0.48	0.46	0.49	0.41
SVD	Infomax	0.51	0.48	0.35	0.38	0.43	0.42	0.48	0.46	0.49	0.43
SVD	JADE	0.53	0.43	0.40	0.37	0.44	0.38	0.50	0.47	0.53	0.40
SVD	ProDenICA	0.49	0.44	0.36	0.48	0.44	0.45	0.41	0.47	0.48	0.41
FastICA	Infomax	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FastICA	JADE	0.96	0.98	0.98	0.99	0.99	0.99	0.98	0.99	0.94	0.98
FastICA	ProDenICA	0.99	0.97	0.98	0.83	0.99	0.99	0.98	0.98	0.96	0.98
Infomax	JADE	0.97	0.97	0.98	0.99	1.00	0.99	0.98	0.99	0.94	0.98
Infomax	ProDenICA	0.99	0.97	0.98	0.85	0.99	0.99	0.99	0.98	0.97	0.98
JADE	ProDenICA	0.96	0.97	0.95	0.83	0.99	0.97	0.96	0.99	0.96	0.96

Method1	Method2	IC 11	IC 12	IC 13	IC 14	IC 15	IC 16	IC 17	IC 18	IC 19	IC 20
SVD	FastICA	0.51	0.61	0.51	0.35	0.39	0.27	0.71	0.59	0.36	0.46
SVD	Infomax	0.51	0.60	0.52	0.32	0.40	0.27	0.74	0.59	0.36	0.48
SVD	JADE	0.55	0.67	0.21	0.31	0.26	0.27	0.70	0.70	0.42	0.25
SVD	ProDenICA	0.42	0.61	0.23	0.44	0.33	0.17	0.79	0.40	0.32	0.18
FastICA	Infomax	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FastICA	JADE	0.99	0.95	0.63	0.97	0.92	0.80	0.95	0.96	0.78	0.74
FastICA	ProDenICA	0.82	0.93	0.69	0.95	0.94	0.89	0.90	0.89	0.89	0.69
Infomax	JADE	0.98	0.96	0.60	0.97	0.92	0.80	0.96	0.96	0.78	0.74
Infomax	ProDenICA	0.80	0.93	0.66	0.94	0.94	0.87	0.90	0.89	0.90	0.69
JADE	ProDenICA	0.83	0.94	0.87	0.96	0.94	0.83	0.85	0.85	0.74	0.95

erate correlations, e.g., less than 0.80, were sometimes associated with large differences. For each IC, we used thresholding and retained 2.5% of voxels corresponding to the most positive values. We visually associated our ICs with networks from Damoiseaux et al. (2006) and present images for selected ICs (Figure A.3). IC 13 has strong lateralization in FastICA and Infomax but is nearly symmetric in JADE and ProDenICA. IC 20 appears to contain areas associated with memory and has strong lateralization in all

methods, but FastICA and Infomax suggest a different spatial configuration than ProDenICA and JADE. ICs 13 and 20 were previously noted to be sensitive to initialization (Section 1.4.2). These networks may be ignored in fMRI studies that use Icasso despite the fact that they do not appear to be artifactual. Parts of the visual cortex are contained in IC4, in which all correlations were greater than 0.80 and the methods look similar, although ProDenICA shows some deviations. IC 3 contains parts of the default network, an area associated with day-dreaming that is often examined in rs-fMRI studies, and the spatial configuration was similar across methods.

We also estimated ICs from a single individual randomly chosen from the ADHD-200 Data Sample. The spatial configuration in individual ICs is less pronounced than in the corresponding group ICs (Figure A.4). The default network (IC 3) in the individual IC is very similar to the group IC, and similarities between IC 4 and IC 20 are also apparent, while IC 13 did not appear to be recovered in this individual.

1.5 Discussion

There is a collaborative effort to share rs-fMRI data from multiple sites in order to improve sample sizes, as in the 1,000 Functional Connectomes Project, the ADHD-200 Sample, and the Autism Brain Imaging Dataset (ABIDE). Thus, there is an urgent need to evaluate whether widely used ICA methods effectively recover resting-state networks, or whether more robust, but typically computationally more expensive, methods produce different results. We have applied a semiparametric method, ProDenICA, to an analysis of rs-fMRI data and demonstrated that multiple initial values are necessary to identify the argmax. In contrast to other fMRI studies, we applied the Hungarian algorithm to match ICs from multiples estimates, and thereby gained novel insights into how some

brain networks are more sensitive to initial values than others, and how some brain network estimates varied little by ICA method while others differ. Given the results from simulations and the fact that IC distributions are rarely, if ever, known in practice, we suggest the use of ICA methods that are effective for a wide range of IC distributions and methods wherein the argmax estimates from multiple initializations correspond to the best estimate. Thus, we suggest ProDenICA be used over FastICA, Infomax, or JADE.

The few studies that considered the impact of starting values on ICA estimation suggested that spurious optima were rarely a problem or excluded ICs that were sensitive to initial values from further analyses; however, we found that ICs that were not sensitive to initialization were the exception and not the rule. In an application of ICA to signal processing, Tichavsky et al. (2005) claimed that approximately 1-100 cases in 10,000 initializations produced estimates from spurious stationary points, and that these cases could be recognized by extremely low signal to interference ratios. In our simulations, local optima were nearly always problematic for twenty components (Figure 1.2). Furthermore, in our fMRI study, spurious optima were found in 40% of initializations for FastICA, 52.5% for Infomax, and 46% for ProDenICA (Figure 1.3).

We argue that evaluating a modest number of randomly chosen initial values and comparing the values of their objective functions is effective and computationally practicable. Tichavsky et al. (2005) proposed a method that imitates a global search for the argmax for a single starting value, although there is no guarantee that it converges to the global maximum. Alternatively, Icasso assumes that cluster centroids accurately characterize ICs. Using cluster centroids produces two sources of error in IC estimates: potential mismatches due to matching via clustering, and error due to the use of cluster centroids instead of the argmax. Furthermore, when multiple estimates of an IC do

not tightly cluster, the IC is typically discarded. Consequently, biologically relevant networks may be ignored simply because their local optima are very different from the argmax. In task-based fMRI, Guo (2011) and Beckmann and Smith (2005) suggest using normal mixtures to model activated and inactivated voxels, but Figures 1k and 1j indicate that FastICA has spurious optima for certain mixtures. Thus, in some cases, biological networks may be ignored in FastICA studies owing to multiple optima that in turn correspond to diffuse clusters.

Moreover, some biological networks may be mis-characterized owing to the poor performance of FastICA and Infomax in recovering some IC distributions, whereas ProDenICA is more robust to IC distributions. For rs-fMRI, the differences between methods were relatively small according to our similarity measures (Table A.4), although visual inspection suggests substantive differences (Figure A.3). In our simulations with two components and very large sample sizes (V = 131,072), Infomax failed for most mixture distributions, and FastICA failed to have a global and/or local maximum at the true unmixing matrix for some asymmetric mixture distributions (Figure 1.1). In contrast, the argmax for ProDenICA with two components corresponded to the true unmixing matrix for all simulated distributions. In simulations with 5, 10, and 20 components, FastICA and Infomax suffered from two problems: oftentimes, an empirical oracle existed that was closer to the true unmixing matrix than the argmax, and secondly, this empirical oracle was inaccurate. JADE was also inaccurate. These issues were resolved in ProDenICA, where the empirical oracle usually corresponded to the argmax, and the argmax was close to the true unmixing matrix (Figure 1.2). These results suggest the difference between methods may be larger in task-based fMRI where normal mixtures model activated/inactive voxels than observed in the resting-state networks.

One approach to examining brain functioning from fMRI studies is to compare mix-

ing matrices between groups, which is often done by assuming a tensor structure that decomposes sources of group variation and sources of individual variation (Beckmann and Smith, 2005; Guo, 2011). In our application, the use of multi-site data with differing numbers of time points precludes the use of a tensor group structure. Here, we focused on the spatial activation patterns rather than the individual and/or group time courses because an examination of mixing matrices of varying dimensions is not trivial. Future research should investigate methods to compare groups where individuals have varying numbers of time points. For example, converting the temporal patterns of activation (columns of $\mathbf{M}^{(m)}$ in (1.8)) to the spectral domain may facilitate an examination of the pathophysiology of diseases.

We conclude that the performance of methods differed dramatically in simulations, and the IC estimates in our fMRI application exhibited variability for some, but not all, ICs. Thus, ProDenICA may improve estimates of ICs in fMRI. Additionally, multiple initial values were essential for identifying the argmax in FastICA, Infomax, and ProDenICA.

CHAPTER 2

LIKELIHOOD COMPONENT ANALYSIS

2.1 Introduction

Transformations that maximize non-Gaussianity play a prominent role in many applications including separating audio recordings in signal processing (Bell and Sejnowski, 1995), denoising in image processing (Hyvärinen et al., 1999), face recognition in computer learning (Bartlett et al., 2002), artifact removal in electrophysiology data (Delorme et al., 2007), and estimating brain networks in cognitive neuroscience (Beckmann, 2012). We propose a novel approach for modeling non-Gaussian signals and Gaussian noise that we call likelihood component analysis (LCA). Consider a sample $\mathbf{x}_1, \ldots, \mathbf{x}_V$ from the LCA model:

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{M}_\mathbf{S}\mathbf{S} + \mathbf{E} \tag{2.1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^T$ is constant; $\mathbf{S} \in \mathbb{R}^Q$ is a vector of mutually independent non-Gaussian random variables with $Q < T$; $\mathbf{M}_\mathbf{S} \in \mathbb{R}^{T \times Q}$ is a fixed rank-Q mixing matrix; and $\mathbf{E}$ is a degenerate multivariate normal random vector with a rank $T - Q$ covariance matrix. Our goal is to estimate $\mathbf{M}_\mathbf{S}$ and the realizations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_V$ of $\mathbf{S}$, which we call latent components (LCs). We apply our method to network estimation and artifact detection in a state-of-the-art functional magnetic resonance imaging (fMRI) dataset with hundreds of thousands of observations and hundreds of variables. We will demonstrate that estimation of the proposed model can allow the discovery of non-Gaussian signals discarded by current methods.

Classic independent component analysis (ICA) and principal component analysis followed by ICA (hereafter, PCA-ICA) are arguably the most commonly used models

for extracting non-Gaussian signals. Unlike (2.1), the classic ICA model assumes $\mathbf{M_S}$ is square (Hyvärinen and Oja, 2000) and $\mathbf{E} = \mathbf{0}$. In practice, PCA is applied prior to classic ICA to meet the assumption of square mixing and to reduce computational costs (Hyvärinen et al., 2001). In this study, we demonstrate that removing the smallest principal components (PCs) can discard the relevant signal (see also Green et al. 2002). For an intuition as to why this is true, first decompose $\mathbf{E}$ into a $T \times (T - Q)$ matrix with orthonormal rows and a vector of $T - Q$ independent normal random variables: $\mathbf{E} = \mathbf{M_N N}$. Now consider the special case where (1) we constrain $\mathbf{M_S}$ to have orthonormal rows; (2) $\mathbf{M_S}$ is orthogonal to $\mathbf{M_N}$; and (3) the singular values of $\mathbf{M_S S}$ are larger than the singular values of $\mathbf{E}$. Then the uniqueness of the singular value decomposition (SVD; defined for the random vector $\mathbf{X}$ using the eigenvalue decomposition of the covariance matrix) implies that the principal subspace will contain the non-Gaussian components; in this case, the PCA-ICA and LCA models are equivalent. If we have conditions (1) and (2) but the singular vales of $\mathbf{E}$ are larger than those of $\mathbf{M_S S}$, then PCA completely discards the non-Gaussian components. In the typical situation where $\mathbf{M_S}$ is not orthogonal and where the singular values of $\mathbf{M_S S}$ and $\mathbf{E}$ are not ordered, there will be some overlap between the PCA-ICA model and LCA model, with the amount of overlap decreasing as the noise increases.

One of the most common applications of PCA-ICA is the identification of brain networks and artifacts in neuroimagery (Beckmann, 2012). In fMRI, the blood oxygen level dependence (BOLD) signal is measured across time at thousands of voxels (three-dimensional analogue of a pixel) across the human brain. ICA of fMRI requires dimension reduction via PCA prior to the application of ICA. ICA can be used to 'unmix' the BOLD signal to reveal the underlying functional architecture of the human brain. The existence and importance of these networks has been corroborated by other neuroimaging modalities and by the application of other statistical methods (Sporns,

2011). Additionally, ICA is commonly used for artifact removal in electroencephalography and fMRI. Independent components (ICs) are identified that correspond to physiological noise and/or motion, and accounting for these artifacts can improve subsequent analyses (Griffanti et al., 2014; Delorme et al., 2007). Even though the results from the two-stage PCA-ICA approach have been useful in the applied sciences, a single analysis that uses non-Gaussianity for both dimension reduction and extracting LCs could provide novel insight.

In this paper, we present a method in which dimension reduction and latent variable extraction are achieved simultaneously to uncover features that are not detected using current models. In section 2, we review existing approaches for extracting non-Gaussian signals. In section 3, we define conditions for the identifiability of the LCA model in (2.1) and propose a parametric and a semi-parametric estimator. In section 4, we investigate simulations when the observations of the latent variables are independently and identically distributed. In section 5, we investigate model robustness by applying our method to temporally and spatially structured simulated data. In section 6, we use our method to estimate brain networks that are engaged in a Theory of Mind (ToM) experiment and artifacts from high-resolution fMRI data from the Human Connectome Project. In section 7, we present our conclusions and discuss avenues for future research.

2.2 Review of alternatives to classic ICA and PCA-ICA

As an alternative to classic ICA, the noisy ICA model posits that the number of noise components is equal to the dimension of the data and typically assumes isotropic noise: $\mathbf{E} \sim N(0, \sigma^2 \mathbf{I}_T)$, where $\mathbf{I}_T$ is the $T \times T$ identity matrix. Beckmann and Smith (2004) propose a variant of PCA-ICA as an approximation to the noisy ICA model, where

they estimate the number of ICs and achieve dimension reduction using probabilistic PCA (Tipping and Bishop, 1999). Alternatively, independent factor analysis (IFA) could be used for simultaneous dimension reduction and latent variable estimation wherein the ICs are modeled as Gaussian mixtures (Attias, 1999). Allassonniere and Younes (2012) developed stochastic EM algorithms to estimate the IFA model and proposed a number of plausible parametric methods. Nonetheless, it is difficult to apply IFA to moderately sized datasets because an m^Q-dimensional integral must be approximated at each iteration, where m is the number of Gaussian mixtures and Q is the number of non-Gaussian components (Allassonniere and Younes, 2012). Amato et al. (2010) develop non-parametric density estimators of the component densities in the noise-ICA model but assume $\mathbf{M_S}$ is semi-orthogonal, which is not realistic for our application.

There are a number of other methods that explore structure in multivariate data using non-Gaussianity. Non-Gaussian measures of information such as kurtosis were first explored in projection pursuit algorithms (Huber, 1985). Non-Gaussian component analysis (NGCA) seeks a lower dimensional subspace that contains the non-Gaussian signal using multiple projection pursuit indices or radial basis functions (Kawanabe et al., 2007). However, NGCA does not model latent components, and thus does not lend itself to identifying brain networks and/or artifacts.

2.3 Modeling latent structure

2.3.1 Identifiability

The identifiability of the LCA model can be established as a corollary to the linear structure model described in Kagan et al. (1973). To simplify the exposition, we assume

$\boldsymbol{\mu} = \mathbf{0}$ in (2.1) (in practice, $\boldsymbol{\mu}$ is estimated as the sample mean of $\mathbf{x}_1, \ldots, \mathbf{x}_v$). We call a random vector $\mathbf{X}$ non-unique if there exist two representations $\mathbf{X} = \mathbf{BY}$ and $\mathbf{X} = \mathbf{CZ}$ in which there exists a column of $\mathbf{C}$ that is not proportional to any of the columns of $\mathbf{B}$. Define the equivalence relation for matrices $\mathbf{B} \cong \mathbf{C}$ if $\mathbf{B}$ equals $\mathbf{C}$ up to scaling and permutation. Let $\overset{d}{=}$ denote equal in distribution. We restate Theorem 10.3.9 from Kagan et al. (1973):

Theorem 1. *Define the linear structure model* $\mathbf{X} = \mathbf{M_S S} + \mathbf{E}$*, where* $\mathbf{M_S}$ *has rank* Q*,* $\mathbf{S}$ *is a random vector with mutually independent non-Gaussian components, and* $\mathbf{E}$ *is a multivariate normal random vector that is non-degenerate for some* $R \leq T$*. Then for any other representation* $\mathbf{X} = \mathbf{M}^*_{\mathbf{S}}\mathbf{S}^* + \mathbf{E}^*$ *where* $\mathbf{S}^*$ *are independent non-Gaussian components, we have:* $\mathbf{M}^*_{\mathbf{S}} \cong \mathbf{M_S}$*;* $\mathbf{S}^* \overset{d}{=} \mathbf{S}$ *up to scaling and permutations; and* $\mathbf{E} \overset{d}{=} \mathbf{E}^*$ *with a non-unique structure.*

If $\mathbf{P}$ is a $Q \times Q$ permutation matrix, $\mathbf{D}$ a diagonal matrix with positive entries on the diagonal, and $\mathbf{J}$ a diagonal matrix with diagonal entries equal to either -1 or 1, then the above theorem states that for any $\mathbf{M}^*_{\mathbf{S}}\mathbf{S}^*$ (such that $\mathbf{S}^*$ has independent non-Gaussian components) and $\mathbf{E}^*$ such that $\mathbf{X} = \mathbf{M}^*_{\mathbf{S}}\mathbf{S}^* + \mathbf{E}^*$, then there exists a $\mathbf{P}$, $\mathbf{D}$, and $\mathbf{J}$ such that $\mathbf{M}^*_{\mathbf{S}}\mathbf{PDJ} = \mathbf{M_S}$.

In the theorem above, $\mathbf{E}$ can have any rank less than or equal to T. In the LCA model, we restrict $\mathbf{E}$ to be rank-$(T - Q)$ multivariate normal, which results in a decomposition of the data into a signal subspace and a noise subspace. Let $\mathbf{S} = [\mathbf{S}_1, \ldots, \mathbf{S}_Q]'$. We summarize the assumptions of the LCA model below:

Assumption 1. *Suppose the model in* (2.1)*. We assume*

i. $\mathbf{S}_1, \ldots, \mathbf{S}_Q$ *are mutually independent, non-Gaussian random variables with* $\mathrm{E}\,\mathbf{S} = \mathbf{0}$ *and* $\mathrm{E}\,\mathbf{SS}^T = \mathbf{I}_Q$*.*

ii. $\mathbf{E}$ *has a decomposition* $\mathbf{M_N N}$ *in which* $\mathbf{M_N}$ *is a rank* $T - Q$ *matrix and* $\mathbf{N}$ *is* $(T - Q)$*-dimensional multivariate normal with* $\mathrm{E}\,\mathbf{N} = \mathbf{0}$.

iii. $\mathbf{M_S}$ *is rank-*Q.

Assumption *i* implies that $\mathbf{X}$ has finite second moments. Define $\mathbf{M} = [\mathbf{M_S}\ \mathbf{M_N}]$. Assumption *ii* implies that the error can be represented as $\mathbf{E} = \mathbf{M_N N}$ such that $\mathbf{N}$ is standard multivariate normal. Along with *iii*, this implies that there exists a full-rank matrix $\mathbf{M} = [\mathbf{M_S}\ \mathbf{M_N}]$, which will be important for computational feasibility.

For the purposes of this paper, we will also assume that the LCs are absolutely continuous and denote their densities $f_1, \ldots, f_Q$, although the results in this section hold more generally. We will find it convenient to define the density of $\mathbf{X}$ in terms of a whitening matrix $\mathbf{L}$ and an orthogonal matrix $\mathbf{W}$. Denote the eigenvalue decomposition of the covariance matrix of $\mathbf{X}$ by $\mathbf{\Sigma} = \mathbf{U\Lambda U}'$. Let $\mathbf{L} = \mathbf{U\Lambda}^{-1/2}\mathbf{U}'$ and define $\mathbf{W}$ such that $\mathbf{WL} = \mathbf{M}^{-1}$. Note that $\mathbf{W} \in \mathcal{O}$, where $\mathcal{O}$ is the class of $T \times T$ orthogonal matrices. Let $\mathbf{w}_q'$ denote the qth row of $\mathbf{W}$, and let $\mathbf{W_S}$ denote the first q rows. Let $\phi(x)$ denote the standard normal density. Noting that $|\det \mathbf{W}| = 1$, we have

$$f_{\mathbf{X}}(\mathbf{x}; \mathbf{W}, \mathbf{L}) = \det \mathbf{L} \prod_{q=1}^{Q} f_q\left(\mathbf{w}_q'\mathbf{L}\mathbf{x}\right) \prod_{k=1}^{T-Q} \phi(\mathbf{w}_{k+Q}'\mathbf{L}\mathbf{x}).$$

Corollary 1. *Suppose the model in* (2.1) *and assumptions i–iii. Then the densities* $f_1, \ldots, f_Q$ *are identifiable, and the vectors* $\mathbf{w}_q$ *for* $q = 1, \ldots, Q$ *are identifiable up to sign (exact if the density is asymmetric). Note that the ordering of* f_q *and* $\mathbf{w}_q$ *is not identifiable, nor are the rows* $\mathbf{w}_{k+Q}$ *for* $k = 1, \ldots, T - Q$.

Proof. Using a change of variable $\mathbf{Z} = \mathbf{LX}$, we consider the model $\mathbf{Z} = \mathbf{A_S S} + \mathbf{A_N N}$, such that $\mathbf{W} = \mathbf{A}'$ and $\mathbf{Z}$ has density $\prod_{q=1}^{Q} f_q\left(\mathbf{w}_q'\mathbf{z}\right)$. Consider $\mathbf{R} \in \mathcal{O}$ and densities

$g_1, \ldots, g_T$ such that

$$\prod_{q=1}^{Q} f_q\left(\mathbf{w}_q'\mathbf{z}\right) \prod_{k=1}^{T-Q} \phi(\mathbf{w}_{k+Q}'\mathbf{z}) = \det \mathbf{L} \prod_{q=1}^{T} g_q(r_q'\mathbf{z}).$$

Let $\mathbf{Q} = \mathbf{R}'$. Then there exists $\mathbf{Y}$ with density $\prod_{q=1}^{T} g_q(y_q)$ such that $\mathbf{Z} = \mathbf{QY}$. From Theorem 10.3.3 in Kagan et al. (1973), $\mathbf{QY}$ has the decomposition $\mathbf{Z} = \mathbf{Q}_1\mathbf{Y}_1 + \mathbf{Q}_2\mathbf{Y}_2$ in which $\mathbf{Y}_1$ are independent non-Gaussian and $\mathbf{Y}_2$ are Gaussian. Then from Theorem (1) and the assumption of unit variance, we have that $\mathbf{Y}_1 \overset{d}{=} \mathbf{S}$, and it follows that there exists a permutation of $g_1, \ldots, g_Q$ equal to $f_1, \ldots, f_Q$. Also from (1), we have $\mathbf{Q}_1 \cong \mathbf{A}_\mathbf{S}$. Using obvious notation, note that $\mathbf{R}_\mathbf{S} = (\mathbf{Q}_1'\mathbf{Q}_1)^{-1}\mathbf{Q}_1'$, and similarly, $\mathbf{W}_\mathbf{S} = (\mathbf{A}_\mathbf{S}'\mathbf{A}_\mathbf{S})^{-1}\mathbf{A}_\mathbf{S}'$, and hence, $\mathbf{R}_\mathbf{S} \cong \mathbf{W}_\mathbf{S}$. □

2.3.2 General LCA Estimator

Now let $\mathbf{x}_1, \ldots, \mathbf{x}_V$ be an iid sample of $\mathbf{X}$. Since $\mathrm{E}\,\mathbf{X} = \mathbf{0}$, we will demean the data such that $\sum_{v=1}^{V} \mathbf{x}_v = 0$. Hereafter, we assume that $\sum_{v=1}^{V} \mathbf{x}_v = 0$. Let $\mathbb{R}_+^{T\times T}$ denote the class of $T \times T$ positive definite matrices. Let $\widehat{\mathbf{\Sigma}}$ be the sample covariance matrix of $\mathbf{x}_v$. Consider its eigenvalue decomposition, $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{U}}'$. Then define

$$\widehat{\mathbf{L}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Lambda}}^{-1/2}\widehat{\mathbf{U}}'.$$

Note that $\sum_{v=1}^{V} \mathbf{o}_q'\widehat{\mathbf{L}}\mathbf{x}_v = 0$. Letting $\mathbf{E} = \mathbf{M}_\mathbf{N}\mathbf{N}$ with $\mathbf{N}$ standard normal, we have

$$\sum_{v=1}^{V} \log \phi(\mathbf{o}_q'\widehat{\mathbf{L}}\mathbf{x}_v) = -\frac{V}{2}\left(\log 2\pi + 1\right). \tag{2.2}$$

Let $\mathcal{O}_{Q\times T}$ be the class of $Q \times T$ semi-orthogonal matrices. Then define the estimator,

$$\widehat{\mathbf{W}}_\mathbf{S} = \underset{\mathbf{O}\in\mathcal{O}_{Q\times T}}{\operatorname{argmax}} \quad \sum_{v=1}^{V}\sum_{q=1}^{Q} \log f_q\left(\mathbf{o}_q'\widehat{\mathbf{L}}\mathbf{x}_v\right). \tag{2.3}$$

Observe that the problem of estimating $\mathbf{W}_\mathbf{S}$ is equivalent to the problem of estimating the LCs because $\hat{\mathbf{s}}_v = \widehat{\mathbf{W}}\widehat{\mathbf{L}}\mathbf{x}_v$ for all v. Thus we would like a consistent estimator of $\mathbf{W}_\mathbf{S}$. Towards this, we have the following lemma:

Lemma 1. *Consider a random vector* $\mathbf{Y} \in \mathbb{R}^T$ *with density* $f_{\mathbf{Y}}$ *such that* $\mathrm{E}\,\mathbf{Y} = \mathbf{0}$ *and* $\mathrm{E}\,\mathbf{Y}\mathbf{Y}' = \mathbf{I}_T$. *Then for any* $\mathbf{o}$, $\mathbf{w} \in O_{T\times 1}$, *we have*

$$\mathrm{E}\,\log\phi(\mathbf{o}'\mathbf{Y}) = \mathrm{E}\,\log\phi(\mathbf{w}'\mathbf{Y}).$$

Proof. It suffices to consider the quadratic term of the Gaussian kernel:

$$\mathrm{E}\,(\mathbf{o}'\mathbf{Y})^2 = \int\left(\sum_{t=1}^{T} o_t y_t\right)^2 f_{\mathbf{Y}}(\mathbf{y})d\mathbf{y}.$$

In the expansion of the quadratic, the cross-terms cancel by our covariance assumption on $\mathbf{Y}$. Then the previous expression is

$$= \sum_{t=1}^{T} o_t^2 \int y_t^2 f_{\mathbf{Y}}(\mathbf{y})d\mathbf{y} = 1.$$

□

Theorem 2. *Suppose* $\mathbf{X}$ *follows the LCA model in* (2.1) *with assumptions i–iii and assume the non-Gaussian components have bounded absolutely continuous densities (satisfied by the classes considered below). Additionally assume* $\mathrm{E}\,\mathbf{X} = \mathbf{0}$ *and* $\mathrm{E}\,\mathbf{X}\mathbf{X}' = \mathbf{I}$ *(here,* $\mathbf{W}_{\mathbf{S}}$ *is the first* Q *rows of* $\mathbf{M}^{-1}$*). Given an iid sample* $\mathbf{x}_1, \ldots, \mathbf{x}_V$, $\widehat{\mathbf{W}}_{\mathbf{S}} \to \mathbf{W}_{\mathbf{S}}$ *almost surely on the equivalence class of signed permutations.*

Proof. Note that $O_{Q\times T}$ is compact. We will show the four assumptions in Wald's consistency proof as recast in Pollard (2001) are satisfied. Let $f_{\mathbf{S}}$ denote the joint density of the LCs. First, we show $\mathrm{E}\,\log f_{\mathbf{S}}(\mathbf{O}\mathbf{X}) \leq \mathrm{E}\,\log f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})$ for any $\mathbf{O} \in O_{Q\times T}$ with equality if and only if $\mathbf{O} \cong \mathbf{W}_{\mathbf{S}}$. Let $\mathbf{W}_{\mathbf{N}}$ denote rows $T - Q$ to T of $\mathbf{W}$. (That $\mathrm{E}\,\log f_{\mathbf{S}}(\mathbf{O}\mathbf{X}) \leq \mathrm{E}\,\log f_{\mathbf{S}}(\mathbf{W}_{\mathbf{S}}\mathbf{X})$ does not hold trivially can be seen by the following

argument:

$$\begin{aligned}
\mathrm{E}\,\log\frac{f_{\mathbf{S}}(\mathbf{OX})}{f_{\mathbf{S}}(\mathbf{W_S X})} &\leq \log \mathrm{E}\,\frac{f_{\mathbf{S}}(\mathbf{OX})}{f_{\mathbf{S}}(\mathbf{W_S X})} \\
&= \log \int \left\{\frac{f_{\mathbf{S}}(\mathbf{Ox})}{f_{\mathbf{S}}(\mathbf{W_S x})}\right\} \{f_{\mathbf{S}}(\mathbf{W_S x})\phi(\mathbf{W_N x})\}\, d\mathbf{x} \\
&= \log \int f_{\mathbf{S}}(\mathbf{Ox})\phi(\mathbf{W_N x}) d\mathbf{x}.
\end{aligned}$$

We would like this quantity to be equal to zero, in which case we would obtain the desired bound; however, $f_{\mathbf{S}}(\mathbf{Ox})\phi(\mathbf{W_N x})$ is a density if and only if $\mathbf{O}$ is orthogonal to $\mathbf{W_N}$, which is not true in general. Consequently, this quantity could integrate to greater than one, in which case we would have $\mathrm{E}\,\log f_{\mathbf{S}}(\mathbf{OX}) \leq \mathrm{E}\,\log f_{\mathbf{S}}(\mathbf{W_S X}) + \alpha$ for some $\alpha > 0$, and thus our bound is not tight enough.)

Define an orthogonal matrix in $O_{T\times T}$ with $\mathbf{O}$ equal to the first Q rows and $\mathbf{O_N}$ equal to rows $Q+1$ to T. Then

$$\begin{aligned}
\mathrm{E}\,\log\frac{f_{\mathbf{S}}(\mathbf{OX})}{f_{\mathbf{S}}(\mathbf{W_S X})} &= \mathrm{E}\,\log\frac{f_{\mathbf{S}}(\mathbf{OX})\phi(\mathbf{O_N X})}{f_{\mathbf{S}}(\mathbf{W_S X})\phi(\mathbf{O_N X})} \\
&= \mathrm{E}\,\log\frac{f_{\mathbf{S}}(\mathbf{OX})\phi(\mathbf{O_N X})}{f_{\mathbf{S}}(\mathbf{W_S X})\phi(\mathbf{W_N X})},
\end{aligned}$$

where the second line follows from Lemma 1. Then

$$\begin{aligned}
\mathrm{E}\,\log\frac{f_{\mathbf{S}}(\mathbf{OX})\phi(\mathbf{O_N X})}{f_{\mathbf{S}}(\mathbf{W_S X})\phi(\mathbf{W_N X})} &\leq \log \mathrm{E}\,\frac{f_{\mathbf{S}}(\mathbf{OX})\phi(\mathbf{O_N X})}{f_{\mathbf{S}}(\mathbf{W_S X})\phi(\mathbf{W_N X})} \\
&= \log \int f_{\mathbf{S}}(\mathbf{Ox})\phi(\mathbf{O_N x}) d\mathbf{x} \\
&= 0,
\end{aligned}$$

which holds with equality if and only if $f_{\mathbf{S}}(\mathbf{Ox})\phi(\mathbf{O_N x}) = f_{\mathbf{S}}(\mathbf{W_S x})\phi(\mathbf{W_N x})$, where the only if direction is a consequence of absolute continuity. Now suppose equality holds and let $\mathbf{Y}$ be a random variable with density $f_{\mathbf{S}}(\mathbf{Oy})\phi(\mathbf{O_N y}) = f_{\mathbf{S}}(\mathbf{W_S y})\phi(\mathbf{W_N y})$. Let $\mathbf{O}_+ = [\mathbf{O}', \mathbf{O}_{\mathbf{N}}']'$. Then there exist random variables $\mathbf{R}_+$ and $\mathbf{R}$ such that $\mathbf{Y} = \mathbf{O}_+\mathbf{R}_+$ and $\mathbf{Y} = \mathbf{WR}$. Applying Theorem 1, we have $\mathbf{O} \cong \mathbf{W_S}$. It follows that

$$\mathrm{E}\,\log f_{\mathbf{S}}(\mathbf{OX}) < \mathrm{E}\,\log f_{\mathbf{S}}(\mathbf{W_S X})$$

for all $\mathbf{O} \ncong \mathbf{W_S}$. The other three conditions are satisfied since we assume continuous, bounded densities and our estimator is an M-estimator. □

Since $\widehat{\mathbf{W}}_\mathbf{S}$ is not invertible, we must also define an estimator of $\mathbf{M_S}$:

$$\widehat{\mathbf{M}}_\mathbf{S} = \underset{\mathbf{B}\in\mathbb{R}^{T\times Q}}{\operatorname{argmin}} \sum_{v=1}^{V} \|\mathbf{x}_v - \mathbf{B}\hat{\boldsymbol{s}}_v\|_2^2.$$

Although we assume iid observations in the construction of (2.3), the LCA model is capable of recovering many forms of dependent data, as is also the case in ICA. This will be demonstrated in simulations.

2.3.3 A parametric model: Logis-LCA

In this section, we present a parametric method called Logis-LCA in which the densities of the LCs are assumed to be logistic. The logistic density is used in the Infomax ICA algorithm, where it appears to work well for unmixing audio signals (Bell and Sejnowski, 1995) and brain networks (Correa et al., 2007). Under the constraint of zero mean and unit variance, the logistic density has the form

$$f_\theta(x) = \frac{\exp\left(-\frac{x}{\theta}\right)}{\theta\left\{1+\exp(-\frac{x}{\theta})\right\}^2}. \tag{2.4}$$

with $\theta = \frac{\sqrt{3}}{\pi}$. We define our estimator for some $\widehat{Q} \le T$ such that $\widehat{Q}$ may or may not equal the true number of components Q. Applying (2.4) and (2.2) to (2.3),

$$\ell(\mathbf{O}; \widehat{\mathbf{L}}, \widehat{Q}, \mathbf{z}_1, \ldots, \mathbf{z}_V) =$$
$$V \log\det\widehat{\mathbf{L}} - \frac{V(T-\widehat{Q})}{2}(\log 2\pi + 1) + \sum_{v=1}^{V}\sum_{q=1}^{\widehat{Q}} 2\log\theta(1+e^{-\mathbf{o}_q'\widehat{\mathbf{L}}\mathbf{x}_v/\theta}) \tag{2.5}$$

where $\theta = \sqrt{3}/\pi$. We maximize (2.5) using the symmetric fixed-point ICA algorithm (Hyvarinen, 1999). Unlike other implementations which require square unmixing matrices, we orthogonalized intermediate estimates of $\mathbf{W_S}$ by calculating the SVD and setting the singular values equal to one. The first and second derivatives of the logistic density are easily calculated, which allows us to calculate an approximate Newton step in the fixed-point algorithm. See Supplemental Materials for details of the algorithm.

2.3.4 A semi-parametric model: Spline-LCA

In this section, we use the flexible family of tilted Gaussian densities to model the LCs. Our proposed method is equivalent to ProDenICA (Hastie and Tibshirani, 2003) when $Q = T$. For $Q < T$, it can be shown that the likelihood is similar to the semiparametric likelihood in Blanchard et al. (2006) but with the independence model for the LCs. The independence assumption is important for physically and biologically useful interpretations.

Suppose the LCs have tilted Gaussian distributions of the form $\phi(x)e^{g(x)}$. Define the log-likelihood:

$$\begin{aligned}&\ell(\mathbf{O}, g_1, \ldots, g_{\widehat{Q}}; \widehat{\mathbf{L}}, \widehat{Q}, \mathbf{x}_1, \ldots, \mathbf{x}_V) \\ &= \sum_{v=1}^{V} \left[\sum_{q=1}^{\widehat{Q}} \left\{ \log \phi(\mathbf{o}_q' \widehat{\mathbf{L}} \mathbf{x}_v) + g_q(\mathbf{o}_q' \widehat{\mathbf{L}} \mathbf{x}_v) \right\} + \sum_{k=1}^{T-\widehat{Q}} \log \phi(\mathbf{o}_{k+\widehat{Q}}' \widehat{\mathbf{L}} \mathbf{x}_v) \right].\end{aligned}$$

This likelihood does not have an upper bound, so we define a penalized likelihood:

$$\begin{aligned}\ell(\mathbf{O}, g_1, \ldots, g_{\widehat{Q}}; \widehat{\mathbf{L}}, \widehat{Q}, \mathbf{x}_1, \ldots, \mathbf{x}_V) = &- \sum_{q=1}^{\widehat{Q}} \lambda_q \int \{g_q''(x)\}^2 dx - \int \phi(x) e^{g_q(x)} dx \qquad (2.6) \\ &+ \frac{1}{V} \sum_{v=1}^{V} \sum_{q=1}^{\widehat{Q}} g_q(\mathbf{o}_q' \widehat{\mathbf{L}} \mathbf{x}_v) - \frac{T}{2}(\log 2\pi + 1),\end{aligned}$$

where we have used (2.2) to simplify the Gaussian components.

Consider the problem of estimating $g(x)$ for fixed $\mathbf{O}$.

Proposition 1. *Let G be the class of all cubic splines $g : \mathbb{R} \to \mathbb{R}$. Consider the argmax of (2.6) for $g_q \in G$. Then (i) $\int \phi(x) e^{g_q(x)} dx = 1$ and (ii) $\int x\phi(x) e^{g_q(x)} dx = 0$ for each q.*

Proof. It suffices to consider the case $\widehat{Q} = 1$. Let $\mathbf{o}_1$ be given and define $s_v = \mathbf{o}_1' \widehat{\mathbf{L}} \mathbf{x}_v$. Define the class of functions $H = \{h : \mathbb{R} \to \mathbb{R},\ h(x) = \theta_0 + \theta_1 x,\ \theta_0, \theta_1 \in \mathbb{R}\}$, and note that H is in the null space of the penalty $\lambda \int \{g''(x)\}^2 dx$. Let $J = \{j \in G :\ \langle h, j \rangle = 0\ \forall\ h \in H\}$. Then $G = H \oplus J$, where $\oplus$ denotes the direct sum. Now let g be the argmax of (2.6) for $g \in G$. Then $g(x) = h(x) + j(x)$ for some $h \in H,\ j \in J$. Then we have

$$\frac{\partial \ell(g)}{\partial \theta_0} = 1 - \int \phi(x) e^{g(x)} dx,$$

from which it follows that $\phi(x) e^{g(x)}$ is a density. Next,

$$\frac{\partial \ell(g)}{\partial \theta_1} = \frac{1}{V} \sum_{v=1}^{V} s_v - \int x\phi(x) e^{g(x)} dx,$$

where we have applied Leibnitz's rule to interchange differentiation with respect to θ_1 and integration with respect to x since $\phi(x) e^{g(x)}$ and $x\phi(x) e^{g(x)} dx$ are continuous on $\mathbb{R}$. Then it follows that $\mathrm{E}\, S = 0$ for S with density $\phi(x) e^{g(x)}$. □

The description of ProDenICA in Hastie and Tibshirani (2003) is in terms of natural quartic splines, in which case the null space of the penalty includes functions of the form $\theta_0 + \theta_1 x + \theta_2 x^2$. Then the above argument can be extended to show that $\mathrm{Var}\, S = 1$, although Hastie and Tibshirani (2003) use cubic B-splines in practice. We standardize the input to have unit sample variance and let G be the class of cubic B-splines. Our estimation follows Hastie and Tibshirani (2003) with modified orthogonalization. We use eight effective degrees of freedom with 100 bins.

2.3.5 A sign and permutation invariant measure for non-square matrices

To assess the accuracy of our estimates and/or compare multiple estimates, we need a discrepancy measure that is invariant on the equivalence class of signed permutation matrices, and we would like a measure that can apply to matrices of differing dimensions when $\widehat{Q}$ may not equal Q. We cannot use the Amari or the minimum distance (Ilmonen et al., 2010) measures because $\mathbf{M_S}$ is non-square. We propose a novel measure of dissimilarity that uses the Hungarian algorithm to match rows of the unmixing matrix as in Risk et al. (2014) but applies to non-square unmixing. We also generalize the measure to apply to matrices that may have a different number of rows, in which case the measure only compares matching rows.

Consider $\mathbf{M}_1 \in \mathbb{R}^{T\times Q}$ and $\mathbf{M}_2 \in \mathbb{R}^{T\times R}$ with $Q \leq R$. Let $\mathcal{P}_\pm$ be the class of $R \times Q$ signed permutation matrices, which results in a subset of Q (permuted) columns of $\mathbf{M}_2$ for $Q < R$. Define the permutation-invariant mean-squared error:

$$PMSE(\mathbf{M}_1, \mathbf{M}_2) = \min_{\mathbf{P}_\pm \in \mathcal{P}_\pm} \|\mathbf{M}_1 - \mathbf{M}_2\mathbf{P}_\pm\|_F^2, \tag{2.7}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\mathbf{P}_\pm$ is found using the Hungarian algorithm. In practice, we also standardize the columns of $\mathbf{M}_1$ and $\mathbf{M}_2$ to have unit norm, and thus the measure is scale invariant. Another advantage of this measure is that it can be used to compare independent components directly. If $\mathbf{S}_1$ is a $Q \times V$ matrix with rows corresponding to independent components (i.e., each column is a sample of the latent vector in $\mathbb{R}^Q$), and if $\mathbf{S}_2$ is $R \times V$, then we define their discrepancy as $PMSE(\mathbf{S}_1', \mathbf{S}_2')$.

2.4 Simulations examining distributional and noise-rank assumptions

In this section, we simulate the LCA model and the noisy ICA model under a variety of source distributions in which the components are iid as well as a scenario in which the sources are sparse images. We compare (1) deflationary fastICA with the log cosh nonlinearity (D-FastICA), where the deflation option estimates components one-by-one such that the algorithm is considered a projection pursuit method (Hyvärinen and Oja, 2000); (2) two-class IFA with isotropic noise (IFA); (3) PCA followed by Infomax (P-Infomax); (4) PCA followed by ProDenICA (P-ProDenICA) (4) Logis-LCA; and (5) Spline-LCA. We evaluate the robustness of these methods with respect to assumptions on the rank of the noise components, distribution of the components, and the signal-to-noise ratio (SNR), as described below.

We fit D-FastICA using the 'deflation' option in the fastICA function with T components from the fastICA R package (Marchini et al., 2010) and select the first Q components. We fit P-Infomax using our own implementation of the Infomax algorithm. We fit P-ProDenICA using the ProDenICA function from the R package of that name (Hastie and Tibshirani, 2010). Note that these methods can provide an estimate of $\mathbf{S}$ but not the mixing matrix, which we estimated as the coefficients from multivariate regression with the data as the response matrix and the estimated components as covariates. We fit the IFA model with two-class mixtures of normals using our own implementation in which $\mathbf{M_S}$ was estimated by maximizing the log likelihood using a numerical optimizer, and the ICs were estimated based on their conditional means (e.g., Amato et al. 2010).

In this paper, we define the SNR as the ratio of the variance from the mixed non-Gaussian components to the variance from the noise components. Formally, consider

the non-zero eigenvalues $\lambda_1, \ldots, \lambda_Q$ from the covariance matrix of $\mathbf{M_S}\boldsymbol{s}_v$ or $\mathbf{M}\boldsymbol{s}_v$ for the LCA or noisy ICA models, respectively. For the LCA model, define $\boldsymbol{\epsilon}_v = \mathbf{M_N}\mathbf{n}_v$ and let $\lambda_{\epsilon_1}, \ldots, \lambda_{\epsilon_{T-Q}}$ denote the eigenvalues from the EVD of the covariance of $\boldsymbol{\epsilon}_v$. Similarly define $\lambda_{\epsilon_1}, \ldots, \lambda_{\epsilon_T}$ for the noisy ICA model. Then,

$$SNR = \frac{\sum \lambda_q}{\sum \lambda_{\epsilon_t}}. \tag{2.8}$$

Let $\mathbf{m}_t'$ denote the tth row of $\mathbf{M}$. For centered data, the sample analogue of (2.8) can be calculated as

$$snr = \frac{\sum_{t=1}^{T}\sum_{v=1}^{V}(\mathbf{m}_t'\boldsymbol{s}_v)^2}{\sum_{t=1}^{T}\sum_{v=1}^{V}\epsilon_{vt}^2}.$$

2.4.1 Simulation Design

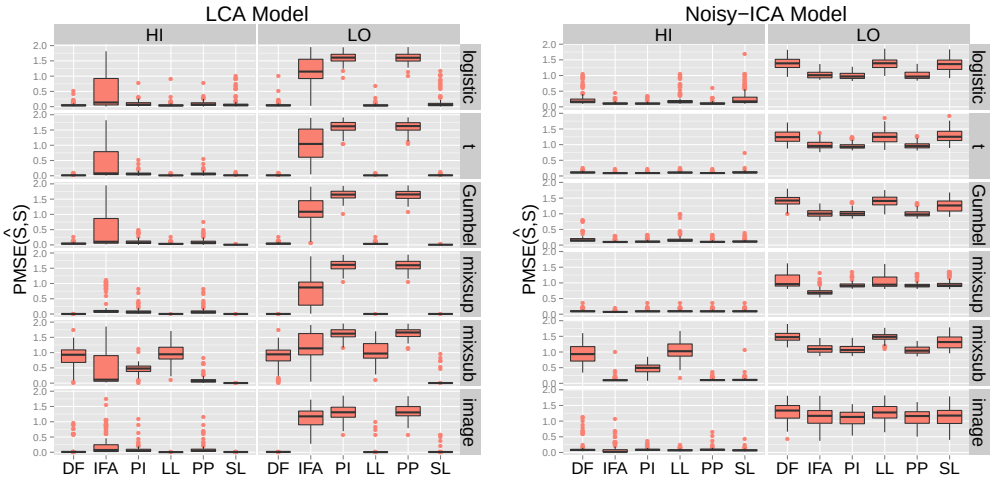
We applied a three-way full factorial design in which data were generated according to (1) the LCA model with rank-$(T - Q)$ noise or the noisy ICA model with rank-T noise; (2) a high or a low SNR; and (3) iid observations from a logistic, t, Gumbel, sub-Gaussian mixture of normals, super-Gaussian mixture of normals, or with values determined by a sparse image, as described below. We generated two signal components for all simulations and used three and five noise components for the LCA model and noisy ICA model, respectively. For the high and low SNR scenarios, the ratio of the variance from the signal components to the variance from the noise components was 5:1 and 1:5. Observations in the noise components were iid isotropic normal except for the sparse image scenario, in which we used the R-package neuRosim (Welvaert et al., 2011) to generate three-dimensional Gaussian random fields with full width at half maximum (FWHM) equal to 6 for each noise component.

The signal components had scale parameter equal to $\sqrt{3}/\pi$ for the logistic, 5 degrees of freedom for the t, and scale parameter equal to $\sqrt{6}/\pi$ for the Gumbel. For the super-

Gaussian mixture of normals, we simulated a two-class model with the first centered at 0 with variance $2/3$ with probability 0.95 and the second centered at 5 with unit variance (excess kurtosis ≈ 9), which is motivated by a brain network with 5% of voxels activated. For the sub-Gaussian mixture of normals, we used the two-class model with the first centered at -1.7 with unit variance and probability 0.75 and the second centered at 1.7 with unit variance and probability equal to 0.25, which is equivalent to distribution 'l' from Hastie and Tibshirani (2003) (excess kurtosis ≈ -0.3). For the sparse image, we used neuRosim to generate an image in which all voxels were iid normal with variance equal to 0.0001 except for a sphere of radius two in which the center was located at $(5, 5, 5)$ with voxel-value equal to one and the exponential decay rate set to 0.5. The second component was similar except the feature was a cube centered at $(7, 7, 7)$ with both radius and exponential decay rate equal to one.

We conducted 112 simulations with 1,000 observations and a random mixing matrix with condition number between one and ten for each combination of factors. Since neither set of orthogonal matrices (PCA-ICA methods) nor semi-orthogonal matrix (LCA methods) is convex, we approximated the argmax by initializing D-FastICA, PCA-Infomax, Logis-LCA, and Spline-LCA from twenty random matrices and selecting the estimate associated with the largest objective function value. For Logis-LCA and Spline-LCA, ten of these twenty initializations were from the principal subspace, i.e., a 5-by-2 semi-orthogonal matrix comprising a 2-by-2 orthogonal matrix and a 2-by-3 matrix of zeros.

Figure 2.1: Boxplots of $PMSE$ for estimated columns of $\mathbf{S}$ where the rank of the noise was $T - Q$ (LCA Model) or T (Noisy-ICA Model) in high SNR ('HI') and low SNR ('LO') scenarios for various latent distributions. 'DF' = D-FastICA; 'IFA' = independent factor analysis; 'PI' = PCA-Infomax; 'LL' = Logis-LCA; 'PP' = PCA-ProDenICA; 'SL' = Spline-LCA.



2.4.2 Results

When the LCA model was true and there was a high SNR, all methods generally produced accurate estimates of $\mathbf{S}$ for the logistic, t, Gumbel, super-Gaussian mixture of normals, and sparse images, but only Spline-LCA was accurate for the sub-Gaussian mixture of normals, and IFA was more variable than other methods for all distributions (Figure 2.1). In these simulations, boxplots examining the accuracy of $\widehat{\mathbf{M}}_{\mathbf{S}}$ showed patterns similar to those found in Figure 2.1 and consequently are not presented.

When the LCA model was true and there was a low SNR, IFA, PCA-Infomax, and PCA-ProDenICA failed to recover the LCs for all distributions, while D-FastICA and Logis-LCA recovered all distributions except for the sub-Gaussian mixture of normals, and Spline-LCA was the most robust to distributional assumptions. Spline-LCA was the only method that recovered the sub-Gaussian mixture.

When the noisy ICA model was true and there was a high SNR, all methods generally produced accurate estimates for the logistic, t, Gumbel, super-Gaussian, and sparse image, although for the logistic distribution, estimates from D-FastICA and Spline-LCA were more variable than the other methods. IFA and Spline-LCA were the only methods that recovered the LCs with the sub-Gaussian distribution.

When the noisy ICA model was true and there was a low SNR, all methods performed poorly, although IFA and PCA-Infomax outperformed the LCA algorithms for all distributions except the sparse image. For the logistic, t, Gumbel, sub-Gaussian mixture, and image, PCA-Infomax and PCA-ProDenICA were slightly more accurate than IFA, although IFA was more accurate for the super-Gaussian mixture of normals.

Overall, LCA methods were robust to the SNR for rank-$(T - Q)$ noise, and performed well in the high SNR scenario for rank-T noise. Additionally, Spline-LCA was most robust to distributional assumptions. In contrast, IFA, PCA-Infomax, and PCA-ProDenICA performed poorly in the low SNR scenario for both the rank-$(T - Q)$ and rank-T noise.

2.5 Simulations examining spatio-temporal networks

In this section, we examine the ability of D-FastICA, PCA-Infomax, Logis-LCA, and Spline-LCA to recover simulated networks whose loadings vary deterministically with time in the presence of spatially and temporally correlated noise, where the simulations resemble the structure found in task-based fMRI, and we examine the effect of $\widehat{Q} \neq Q$ on network recovery. In this way, we assess whether the LCA algorithm can recover brain networks and their temporal loadings from spatiotemporal neuroimagery. We did not include IFA in these simulations because it was difficult to estimate the mixing

matrix when T was relatively large (e.g., $T = 50$). Additionally, IFA and PCA-Infomax produced similar results for most distributions in the previous simulations, and for rank-$(T - Q)$ noise, PCA-Infomax was more accurate than IFA in the high SNR scenario, and IFA and PCA-Infomax performed similarly in the low-SNR image scenario. Hence, our previous simulations suggest there would be little insight gained from including IFA.

2.5.1 Simulation Design

We simulated three networks mixed across fifty time units. The networks were 33×33 images where "active" pixels were in the shape of a "1", "2 2", or "3 3 3" with values between 0.5 and 1 and "inactive" pixels were mean zero iid normal with variance equal to 0.0001 (see Figure 2.2). Let $\mathbf{m}_q$ denote the qth row of $\mathbf{M_S}$ (or qth row of $\mathbf{M}$ for noisy ICA simulations). To simulate the temporal activation patterns of brain networks, we used neuRosim to convolve the canonical hemodynamic response function (HRF) with a block-design with onsets at $\{1, 20.6\}$, $\{10.8, 40.2\}$, and $\{10.8, 30.4\}$ for $\mathbf{m}_1$, $\mathbf{m}_2$, and $\mathbf{m}_3$, respectively, and duration equal to 5 time units.

In the LCA scenario, noise components were generated as forty-seven independent 33×33 Gaussian random fields with FWHM=6. Temporal noise structure was introduced via the mixing matrix, in which each column of $\mathbf{M_N}$ corresponded to an AR(1) process simulated for fifty time units with AR coefficient equal to 0.47 and unit variance, where the AR coefficient was chosen based on a preliminary analysis of the fMRI data analyzed in Section 2.6. Additionally, noise components were scaled such that the SNR was 0.4, which approximately equals the SNR estimated in Section 2.6. In the noisy ICA scenario, a 33×33 Gaussian random field with FWHM=6 was simulated for $t = 1$ and then noise components were defined recursively for $t = 2, \ldots, 50$ to be equal to 0.47

times the noise at time $t-1$ plus a simulation from an independent Gaussian random field with FWHM=6.

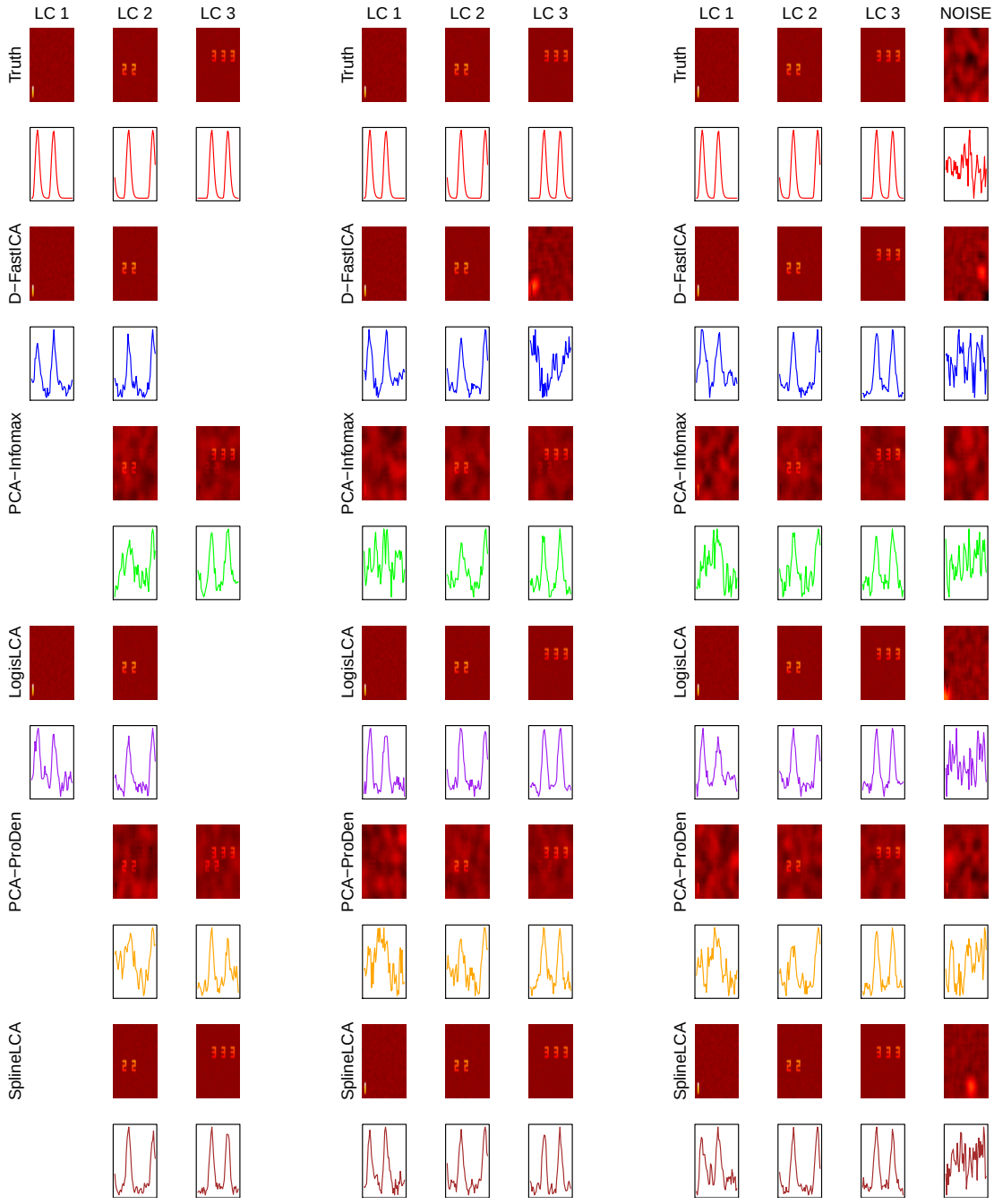
We conducted 111 simulations with $\widehat{Q} = 2, 3$ or 4 and initialized all algorithms from twenty random mixing matrices for each simulation and each $\widehat{Q}$. For Logis-LCA and Spline-LCA, ten of the twenty initializations were from random matrices in the principal subspace, as in Section 2.4.1.

2.5.2 Results

By inspecting the images and loadings associated with the median $PMSE(\hat{\mathbf{S}}, \mathbf{S})$ for each method in the LCA scenario, we see that D-FastICA recovers a spurious component $\widehat{Q} = 3$, PCA-Infomax and PCA-ProDenICA generally fail to unmix features; and Logis-LCA and Spline-LCA are highly accurate (Figure 2.2). It is notable that estimates from PCA-Infomax and PCA-ProDenICA were sensitive to the choice of $\widehat{Q}$, as when $\widehat{Q} < Q$, an estimated latent component resembled a union of components two and three. In PCA-ProDenICA, the loadings for the estimated component were highly correlated with component three (R=0.75), which mistakenly suggests components three and two are functionally connected. For $\widehat{Q} = 3$, the features in the estimated component one are faintly visible in PCA-Infomax whereas component one was not recovered by PCA-ProDenICA. In contrast, Logis-LCA and Spline-LCA clearly separated components for all $\widehat{Q}$, such that when $\widehat{Q} < 3$, the recovered components were accurate estimates of a subset of the true components.

For the noisy-ICA scenario, the features recovered by Logis-LCA most closely resembled the truth (Figure 2.2). Features from component two were again faintly visible in component three for $\widehat{Q} = 2$ in both PCA-Infomax and PCA-ProDenICA, again in-

Figure 2.2: Network recovery from the LCA scenario with $Q = 3$ for $\widehat{Q} = 2$ (first three columns), $\widehat{Q} = 3$ (columns 4-6), or $\widehat{Q} = 4$ (columns 7-10). Images depict LCs and time-series plots depict the loadings corresponding to the median $PMSE(\widehat{\mathbf{S}}, \mathbf{S})$ from 111 simulations. In the last column, the first two rows correspond to an arbitrary noise component whereas the algorithms attempted to estimate a fourth LC.



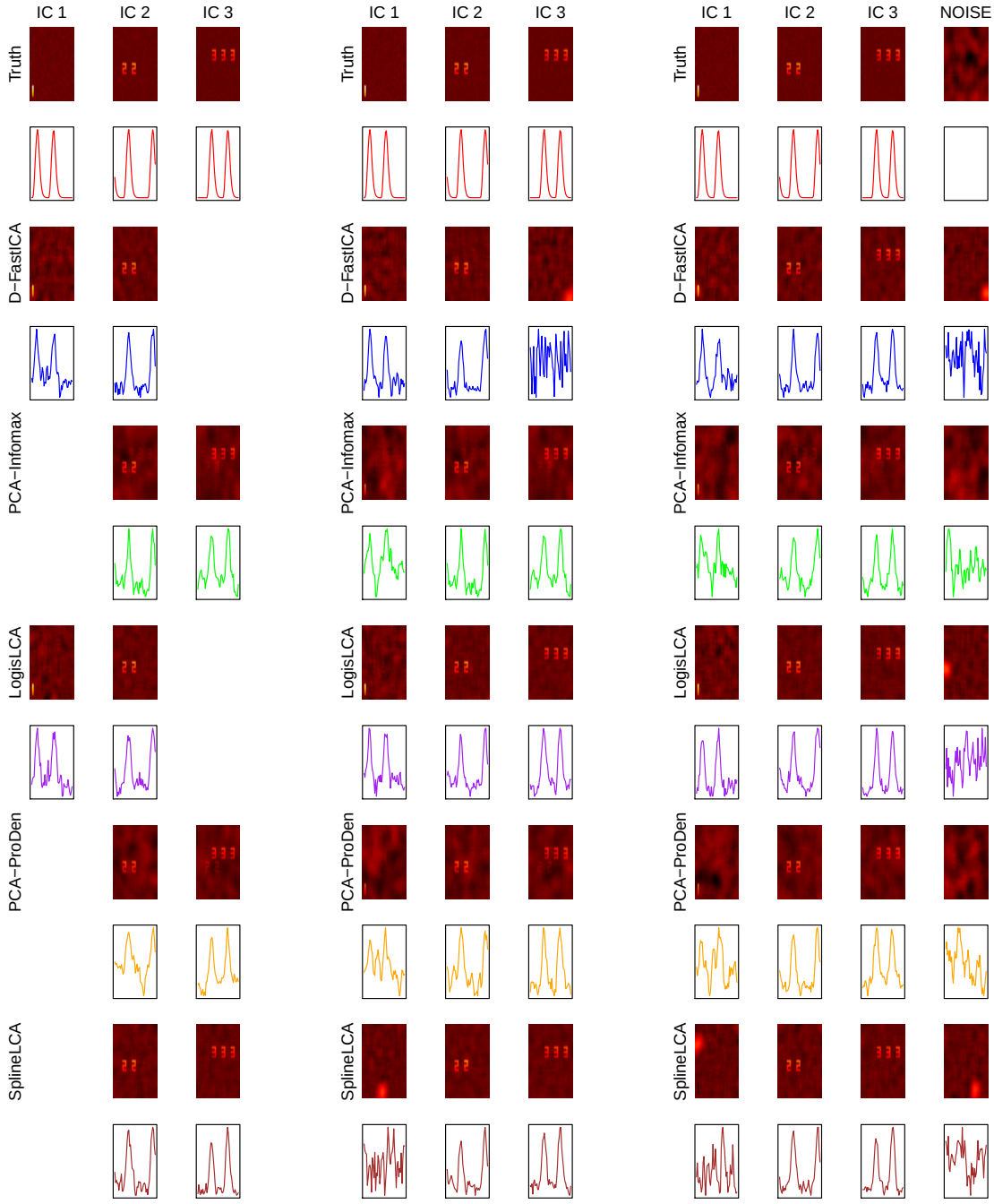
dicating inadequate unmixing of the networks. As seen in the rank-$(T - Q)$ scenario, D-FastICA recovered a spurious component for $\widehat{Q} = 3$, but accurately estimated component three for $\widehat{Q} = 4$. Spline-LCA was sensitive to the assumption on the rank of the noise, as it failed to recover component one, although it was quite accurate for components two and three. Spatial correlations in the noise appear to result in spurious circular features, which were detected in Spline-LCA and D-FastICA. In general, an accurate estimate of component one was associated with a local maxima in Spline-LCA, whereas the spurious component had a higher likelihood.

2.6 Application to fMRI

We applied Spline-LCA to a single subject from the Social Cognition / Theory of Mind (ToM) experiment of the MGH-UCLA Human Connectome Project (HCP; additional information in Supplemental Materials). For details of the experimental paradigm see Barch et al. (2013). We used the minimally pre-processed data (Glasser et al., 2013) from the first session of subject 103414 from the June 5, 2014, data release. The first two volumes were removed to allow for scanner equilibration. Three-dimensional volume data were vectorized and non-brain tissue excluded using the mask provided from the HCP. This resulted in a 230,459 $\times$ 272 data matrix. Each voxel was treated as a replicate with $v = 1, \ldots, V$, which is analogous to 'spatial' ICA of fMRI (Calhoun et al., 2009).

The application of ICA to fMRI commonly assumes that voxels are iid. This assumption is often not made explicit because ICA is usually derived from the perspective of maximizing non-Gaussianity. Since the fixed-point algorithm can also be derived from ML theory where the non-linear function is equivalent to the log likelihood (e.g., Hyvärinen and Oja 2000), summation of the non-linear function over voxels (e.g., Equa-

Figure 2.3: Network recovery from the noisy-ICA scenario with $Q = 3$ for $\widehat{Q} = 2$ (first three columns), 3 (columns 4-6), or 4 (columns 7-10).



tion 12 in Beckmann and Smith 2004) is mathematically equivalent to assuming the voxels are independent. Despite the violation of model assumptions, ICA recovers simulated brain networks and their loadings (Beckmann and Smith, 2004) and has proven useful in constructing models of functional connectivity that are consistent across subjects and image acquisition centers (Biswal et al., 2010). Thus we follow studies using ICA of fMRI and assume iid voxels. Additionally, we mean centered and variance normalized each voxel's time course prior to conducting LCA, since this normalization was suggested for ICA of fMRI (Beckmann and Smith, 2004).

We used the ICA software MELODIC (FSL) to determine the number of components that would be used in an analogous ICA of this dataset. Using this software, thirty components were chosen. We initiated the algorithm from fifty-six randomly generated matrices, twenty-eight of which were in the principal subspace. We selected the estimate corresponding to the largest log likelihood as our estimate of the true argmax. Depending on initialization, the algorithm took between ten minutes and 3.75 hours on a 2666 MHz processor, where 3.75 hours represented initializations that reached the maximum number of iterations, which we conservatively chose to be equal to 300. We also completed an analogous PCA-ProDenICA with thirty components and fifty-six initializations using the R package ProDenICA (Hastie and Tibshirani, 2010).

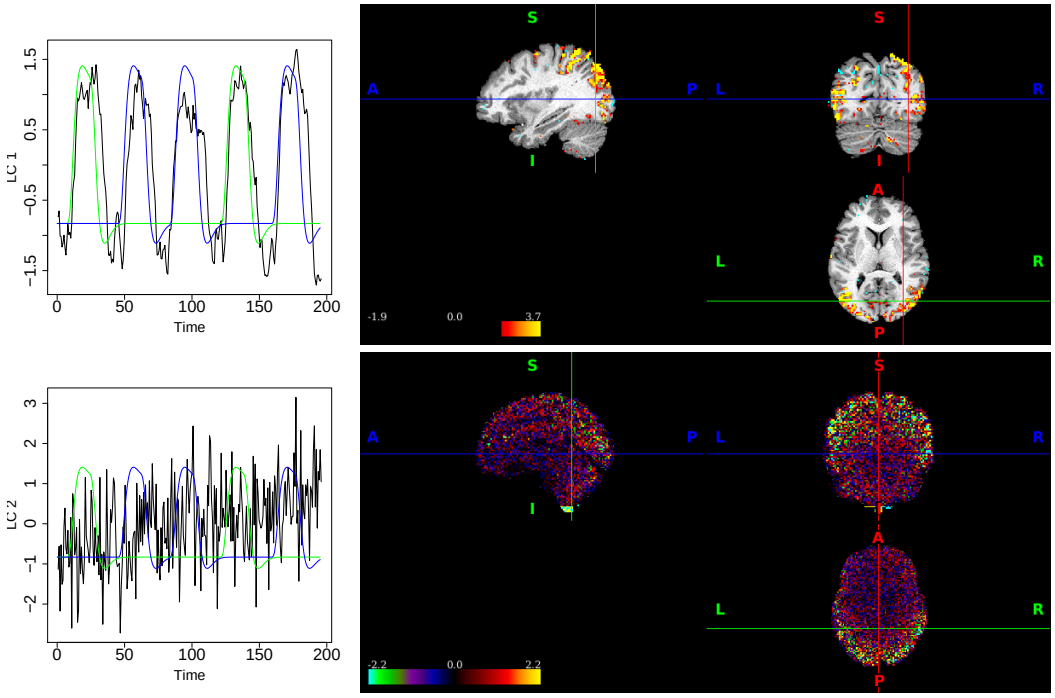
We examined the correlation between the loadings for each component (columns of $\widehat{\mathbf{M}}_{\mathbf{S}}$) to the two covariates 'mentalize' and 'random.' These covariates were generated by convolving each task's onsets and durations with the canonical HRF in SPM8 (Functional Imaging Laboratory). The first component was highly correlated with the mentalize and random tasks (Figure 2.4). This component showed activation primarily in the lateral occipital cortex. A similar component was found using PCA-ProDenICA (not depicted).

We also detected components that were estimated in Spline-LCA but not in PCA-ProDenICA. Eight out of thirty LCs had a correlation less than 0.5 with their matched IC components. In particular, component two in Spline-LCA was not correlated with any of the components in PCA-ProDenICA (max correlation among all ICs = 0.01). This component appears to correspond to an artifact due to motion and possibly other sources of noise. Its time course was correlated with three of the motion parameters from the rigid-body alignment (r = 0.32, 0.32, and 0.42 for the x-transformation, x-rotation, and z-rotation parameters, respectively). Voxels were highly activated in the brainstem, which could be due to movement. Additionally, there was a positive correlation with time (r = 0.44), which could be related to scanner drift. Removing artifacts from fMRI detected using ICA is a popular tool that can increase detection in subsequent mixed-modeling of voxel activation (Tohka et al., 2008). Thus, our detection of a novel artifact represents a potential benefit of LCA over current methodology.

2.7 Discussion

In this study, we propose a model-based method for estimating non-Gaussian latent components in the presence of Gaussian noise that has many applications including signal processing, psychometrics, and computer learning, and we applied the method to identifying brain networks and artifacts from neuroimagery. Simulations indicate that our methods perform well for low SNR when the LCA model is true. When the noisy ICA model is true, our methods perform well in the high SNR scenario, while none of the methods perform well in the low SNR scenario. At the moderately low SNR used in the fMRI simulations (0.47), Logis-LCA and Spline-LCA outperformed competing methods for the LCA model and Logis-LCA outperformed PCA-Infomax for the noisy-ICA model. These results suggest that Logis-LCA and Spline-LCA can

Figure 2.4: Selected brain networks estimated from the HCP ToM data using Spline-LCA. The first row depicts a task-activated component that was highly correlated with the mentalizing (green) and random (blue) tasks (MNI coordinates: 35,-75,8; thresholded $|s_{v1}| \geq 2$); a similar component was found using PCA-ProDenICA (not depicted). The second row appears to be an artifact (MNI: 0,-50,0; unthresholded); this component was not found by PCA-ProDenICA.



be used to reveal structure for a large class of non-Gaussian observations that may be spatially and/or temporally dependent. In our application, we simultaneously achieved dimension reduction and latent variable extraction for large image data ($T = 272$ and $V = 230,459$) and identify an artifact not identified by PCA-ICA.

An interesting advantage of LCA over existing frameworks is its robustness to misspecification of the number of estimated components. This robustness suggests LCA could be used to improve estimates of functional connectivity in fMRI studies. In contrast, estimating the correct number of components in noisy ICA is a pre-requisite to recovering valid components (Section 2.5, see also Allassonniere and Younes 2012).

Beckmann and Smith (2004) explored the use of probabilistic PCA to estimate the number of brain networks prior to ICA in order to avoid model overfitting, which addresses the concern that overfitting may 'split' a single network into multiple networks. However, our simulations suggest that too few components leads to inappropriately aggregated networks in PCA-ICA methods (Figures 2.2 and 2.3). In contrast, the components recovered for $\widehat{Q} \leq Q$ in Logis-LCA across model scenarios (Figures 2.2 and 2.3) and Spline-LCA for the LCA scenario (Figure 2.2) accurately represent functional connectivity. Although we have argued that our framework is robust to misspecification of the number of components, we would like a rigorous method to determine the number of components. For iid data, AIC is an effective method to determine the number of components, but AIC and other model selection criteria are ineffective when observations are positively correlated. A likelihood incorporating spatiotemporal dependencies could be pursued over the iid formulation, and then classic model selection criteria could be used. However, our simulations suggest the iid likelihood accurately recovers dependent data, so the improvements from a spatiotemporal likelihood could be outweighed by the computational costs. Future research should investigate selection criteria for non-iid data.

LCA offers a computationally tractable alternative to one of the most common applications of ICA to fMRI and EEG: artifact detection. Currently, PCA-ICA is used as a pre-processing step to reveal biologically implausible loadings and/or loadings resembling physiological artifacts that can be used to de-noise data for subsequent analyses (Beckmann, 2012). In LCA, these artifacts appear as LCs since they have non-Gaussian distributions. Our detection of the artifact in component two (Figure 2.4) suggests LCA could be used for more powerful denoising methods over traditional PCA-ICA. Artifacts may increase and/or become more problematic when using state-of-the-art data with high-resolution and fast acquisition times, as smaller voxels are associated with

smaller signals, indicating artifact removal is increasingly important (Griffanti et al., 2014). The HCP data represent the highest resolution and fastest acquisition times currently available in fMRI, and thus LCA offers a promising alternative to ICA for artifact detection.

Code implementing Logis-LCA, Spline-LCA, and the PMSE similarity measure is available by request.

CHAPTER 3

SPATIOTEMPORAL MIXED MODELING OF MULTI-SUBJECT FMRI: A RETURN TO NORMALCY

3.1 Introduction

A goal of neuroscience is to map the locations of brain activity that are associated with various thoughts and behaviors. Brain activity can be measured indirectly using functional magnetic resonance imaging (fMRI), which measures the BOLD (blood oxygen level dependent) signal over a grid of voxels (volumetric pixels) across time. A linear mixed model can be used to estimate the relationship between the BOLD signal and the time courses of task stimuli while accounting for subject-specific variation. Although fMRI data contain many spatial and temporal correlations, accounting for these dependencies is difficult due to its large size. Consequently, the primary tool to map regions of activation is a massive univariate analysis in which a separate mixed model is fit to each voxel (Worsley et al., 2002). Estimating a statistical model at each individual voxel requires spatial smoothing. Spatial smoothing serves two main functions: it can increase the power to detect activation by increasing the signal-to-noise ratio, and it increases the overlap of corresponding features. However, this comes at the expense of decreasing the precision with which activation is localized. Moreover, popular approaches to smoothing decrease the information available from technological improvements by reducing the effective resolution (Tabelow et al., 2009).

There is no clear answer to how much data smoothing should be applied. The issue is exacerbated by the fact that inference is sensitive to the amount of smoothing (Mikl et al., 2008). We view the amount of smoothing as an unknown parameter that attempts to balance the increase in overlapping features between subjects with the precision of

localization. In this paper, we propose a novel spatiotemporal mixed model for large, multi-subject data. Our contributions are the following. First, we introduce spatial random effects that capture population activation, which leads to automated smoothing. This obviates the need for smoothing to increase the power to detect activated locations, since the amount of power is now determined by the data. Second, we utilize subject-location random effects to allow subject-specific deviations in activation and/or alignment. This obviates the need for smoothing to increase the overlap of features between subjects. Third, we develop a unified model that includes subject- and location-specific autoregressive errors, which contrasts with previous methods that use the output from a first-level analysis. Fourth, we leverage improvements in cortical registration and improvements in parcellation to develop a parcel-specific dependence structure for the cerebral cortex. Fifth, we develop fast estimators of spatial dependence that can be used for whole-brain studies, which improve upon previous multi-subject spatial models (discussed below) that assume a constant correlation between all locations within a region.

Although the most popular methods assume that voxels are independent during model fitting, it is common to account for spatial dependencies post-estimation by applying random field theory (RFT) to the statistical images; however, RFT requires preprocessing with a spatial smoother (Worsley et al., 1996). For a covariate or contrast of interest, a test statistic image can be formed from the univariate models fit to each location, which can be converted to a z-statistic image. The full width at half maximum (FWHM = approximately 2.355 times the standard deviation) of the z-statistic image can be estimated, which is a function of both the intrinsic data smoothness and the smoothing kernel used in pre-processing. Clusters of activated regions are created by thresholding voxel-specific t-statistics. The estimated FWHM is then used to parameterize the null Gaussian field from which a critical value for either the extent of a cluster (an excursion set) or the maximal value of a cluster can be calculated. The statistical power

of RFT inference generally increases when the FWHM used in preprocessing increases. Moreover, a number of approximations used in RFT-based critical values assume a relatively large degree of smoothing. One school of thought recommends choosing a equal to three to four times the voxel resolution (Nichols and Hayasaka, 2003). As previously noted, more smoothing comes at the expense of saying precisely where in the cluster activation occurs. Thus there is a need to develop a model that does not rely upon *ad hoc* spatial smoothing yet allows for voxel-level inference on multiple subjects.

Most studies in spatial and/or spatiotemporal modeling of fMRI use Bayesian approaches, but due to computational feasibility these models are either for a single-subject or use a hybrid approach that does not incorporate temporal autocorrelation. The majority of Bayesian methods have focused on single-subject data, where spatial priors such as Gaussian Markov random fields (GMRF) are used to achieve model-based smoothing (Penny et al., 2005). Variational Bayes approaches scale the analysis to the whole brain by approximating the posterior (Harrison and Green, 2010). In order to analyze multiple subjects, hybrid methods have been developed that utilize subject- and vertex- specific coefficients from a first-level ordinary least squares (OLS) analysis (or GLS, generalized least squares accounting for temporal correlation), and then fit a Bayesian model to the first-level output. Xu et al. (2009) fit a spatial Poisson process to the t-statistics from a first level analysis. Derado et al. (2013) fit a conditionally autoregressive (CAR) model in each brain parcel to capture short-range dependencies and model the correlation between parcels to capture longer-range dependencies. For a review of Bayesian approaches, see Zhang et al. (2015). Although scalable, the multi-subject hybrid approaches do not incorporate the variance from the first-level analysis in their second-level estimators, but rather use the first-level coefficients or t-statistics as the responses in their Bayesian modeling. Another potential drawback is that the marginal properties of spatial autoregressive models are often undesirable. For instance, the variance may

depend on the number of neighbors. Boundary effects can be problematic, particularly in three dimensions with parcellations in which the regions are small to moderately sized.

A number of non-Bayesian spatial models have been developed for neuroimaging data and they appear to increase statistical power, but their application to whole brain fMRI data from multiple subjects is either computationally tractable, but with biologically undesirable dependence structures, or computationally problematic. The most scalable of these methods uses a parcellation of the brain wherein the correlation between residuals within a parcel is constant and parcels are assumed independent (Derado et al., 2010; Bowman, 2005). When analyzing fMRI, these models are again applied to the coefficients from the first-level analysis, which are taken as given.

An alternative to assuming constant correlation within a parcel is to use a covariogram that is a function of distance. Deviations from some population level of activation can be modeled using spatial random effects that compose a Gaussian random field. Bowman (2007) evaluated a number of parametric covariograms and used a functionally based measure of distance, and found empirical support for the exponential covariogram. The model included a repeated-measures component allowing for constant correlation between serial measurements and was applied to positron emission tomography (PET) data comprising 239 voxels from twelve subjects with four time points each. Bernal-Rusiel et al. (2013) applied the model in Bowman (2007) to a surface-based analysis of cortical thickness from hundreds of subjects with one to seven sessions each. Their spatial modeling led to large gains in statistical power while controlling the type-1 error rate. To scale the model to the entire cortex, they classified 149,000 locations to 12,000 independent parcels. They also used either 8 or 15 mm FWHM smoothing, which was determined based on their prior experience. Parameters were estimated using restricted maximum likelihood (REML) with a Fisher-scoring algorithm and GLS. Whereas these

previous studies did not use fMRI data, Kang et al. (2012) transformed fMRI data to the spectral domain to simplify the temporal covariance structure. They used the empirical covariogram to capture spatial dependencies and modeled correlations between regions. However, their model is for single-subject data, the empirical covariogram is not positive definite in general, and there is some subjectivity in selecting which bandwidths to analyze in the spectral domain. Hyun et al. (2014) improve spatial prediction by developing a spatial Gaussian predictive model for multiple subjects that utilizes fixed rank kriging with functional principal components and a spatial autoregressive model. They applied their model to diffusion tensor imaging (DTI) data, which does not include a temporal component. Although these methods offer marked improvements over voxel-wise approaches, the extension to whole brain fMRI data involving hundreds of time points is not trivial.

A potential drawback to using a covariogram to model spatial dependence in neuroimagery is that Euclidean distance may only loosely relate to dependence between voxels. Consider voxels that lie on the cortical surface, which is the gray matter comprising most of the neural cell bodies in the brain. The cortical sheet can be visualized as a deflated balloon with many infoldings. Two voxels on the cortical surface may be adjacent in volume space but on opposite sides of a sulcus (fold), and consequently have little dependence. Bowman (2007) accounted for this possibility by using a functional distance measure based on a separate study of the correlation in BOLD signal between voxels. Alternatively, one could restrict attention to the cortex, and then measure the distance along the cortical surface between two vertices, which may be more related to spatial dependence than distances in volume space. Specifically, voxels from volume fMRI data can be classified as cortical tissue and then registered to the cortical sheet. The BOLD signal is represented at locations, called vertices, on the surface, and the vertices form triangles, which in turn compose a tessellated surface. Surface-based

registration has led to large improvements in the ability to assess localized structural differences between subjects (Fischl et al., 1999). Bernal-Rusiel et al. (2013) used a surface-based representation in their MRI study of cortical thickness, and developed their own method to create a parcellation scheme. But in fMRI, abrupt changes in resting state connectivity (i.e., correlations in BOLD signal) have been observed in the cortex. Gordon et al. (2014) created a parcellation such that resting-state connectivity patterns were homogeneous within each region, and in particular, more homogeneous than alternative parcellations such as Brodmann areas. This suggests a method to nest a geodesic distance-based dependence structure within the Gordon parcellation.

We propose a spatiotemporal model to detect activation in fMRI. As in Bowman (2007) and Bernal-Rusiel et al. (2013), we incorporate subject-specific spatial random effects. Unlike previous studies, we introduce "population" spatial random effects to model deviations from the overall parcel activation. This enables us to achieve model-based smoothing for multi-subject fMRI data. To overcome issues with volume-based Euclidean distances and abrupt changes in spatial dependence, we focus on the cortical surface and use geodesic distances between vertices within the same Gordon parcel, and assume parcels are independent. Unlike previous multi-subject fMRI studies, we model the BOLD signal rather than the output from a first-level analysis. Note that we can not use maximum likelihood or restricted maximum likelihood methods to fit this model due to the enormous size of the covariance matrix, which is non-sparse for each parcel. In our application, the largest parcel covariance matrix is 11 million by 11 million. The overall covariance matrix is 326 million by 326 million. Our model uses reduced-biased estimators of AR parameters to model temporal autocorrelation, as proposed in Worsley et al. (2002), and we derive our own estimators of the spatial dependence parameters. Given these dependence parameters, we derive method of moments (ANOVA-like) estimators of the variance components. We apply the model to

high resolution, state-of-the-art fMRI data from a theory of mind (ToM) experiment in the Human Connectome Project (HCP).

The remainder of this paper is organized as follows. In Section 2, we formalize the massive univariate model most commonly used in fMRI analyses, and we describe the popular two-stage OLS estimator. We also discuss averaging the signal across voxels in a parcel for region-of-interest (ROI) analyses. In Section 3, we propose a spatiotemporal mixed model. In Section 4, we conduct simulations to demonstrate the accuracy of our estimators and compare with the massive univariate and ROI approaches. In Section 5, we apply our model to the right cerebral cortex of the ToM HCP data. In Section 6, we conclude that our spatiotemporal mixed model localizes fMRI activation and automates smoothing.

3.2 The Massive Univariate Mixed Model (MUMM) of fMRI

In this section, we present the commonly used two-level hierarchical model for activation in which each location is modeled independently. This model is sometimes referred to as the group general linear model (GLM) approach (Lindquist, 2008) or as a random effects analysis (Penny et al., 2003). Here, we call this the Massive Univariate Mixed Model (MUMM) of fMRI. This model is generally presented for volumetric modeling and hence locations are called voxels, but here, we conduct surface modeling and hence locations are called vertices.

We suggest that most authors agree upon the general structure for a model of activation at a single vertex, which is what we define as the MUMM. Different papers and competing software propose different estimators, including the "summary statistics approach" (Holmes and Friston, 1998), REML using the EM algorithm (Worsley et al.,

2002), and a "hybrid" GLS and Bayesian approach used in FSL (Beckmann et al., 2003; Woolrich et al., 2004); estimation is described in Section 3.2.3. (These papers tend to disagree on the model for the temporal correlation, but agree on the other variance components, as described below.)

3.2.1 First level (subject effects)

Let $n \in \{1, \ldots, N\}$ denote subject, $v \in \{1, \ldots, V\}$ denote the vertex, and $t \in \{1, \ldots, T\}$ index time. For the covariates, let $q \in \{1, \ldots, Q\}$ index a task or other covariate of interest and let $m \in \{1, \ldots, M\}$ index a nuisance covariate, e.g., covariates to capture scanner drift, heart rate, breathing rate, or motion parameters. A table containing notation used throughout this document is provided in the Appendix (Table C.1). In this document, a "task" is considered to be any condition that is being modeled. For example, subjects complete two tasks in our application. In the first, a subject views shapes interacting with each other. In the second, a subject views shapes moving in random ways.

In fMRI analyses, the design matrices for each vertex are usually equivalent. Since the covariates do not change with space, we let x_{ntq} be the covariate for the qth task for the nth subject at the tth time point for all vertices. (Details regarding x_{ntq} are in the next paragraph.) Let $\mathbf{x}_{nt} = [x_{nt1}, \ldots, x_{ntQ}]'$. Let $\mathbf{z}_{nt} \in \mathbb{R}^M$ denote nuisance variables for the nth subject. Let y_{nvt} denote the BOLD signal at the tth time point for the vth vertex of the nth subject. Let a_{nvt} be the error for the nth subject, vth vertex, and tth time point, and let $\boldsymbol{a}_{nv} = [a_{nv1}, \ldots, a_{nvT}]'$. Let e_{nvq} denote the magnitude of activation to the qth task for the nth subject at the vth vertex, and let $\boldsymbol{e}_{nv} = [e_{nv1}, \ldots, e_{nvQ}]'$. Let γ_{nvm} denote the coefficient for the mth nuisance covariate for subject n and location v, and define

$\boldsymbol{\gamma}_{nv} = [\gamma_{nv1}, \ldots, \gamma_{nvM}]'$. Then define the first-level model:

$$y_{nvt} = \mathbf{x}'_{nt}\boldsymbol{e}_{nv} + \mathbf{z}'_{nt}\boldsymbol{\gamma}_{nv} + a_{nvt} \tag{3.1}$$

with

$$\boldsymbol{a}_{nv} \sim N(\mathbf{0}, \xi^2_{nv}\boldsymbol{\Psi}_{nv}) \tag{3.2}$$

where $\mathbf{0}$ is a vector of length T, $\boldsymbol{\Psi}_{nv}$ captures the correlation between serial observations, and $\boldsymbol{a}_{nv}$ are independent for all $n = 1, \ldots, N$ and $v = 1, \ldots, V$. We will refer to $\boldsymbol{\Psi}_{nv}$ as the error variance.

The covariates of interest are calculated from a model for the shape of the BOLD signal when stimulated, which is called the hemodynamic response function (HRF), convolved with the task onsets and durations. In the simplest case, the same HRF is assumed for all locations. In practice, this assumption can be relaxed by including partial derivatives of the parameters of the HRF evaluated across time, such that coefficients are estimated to allow the HRF to vary while the covariates remain the same at every location. This is examined in detail in Appendix C.2. Let $I_{nq}(t)$ denote the onset and durations (box-car function) for the qth task of the nth subject. Let $h(t)$ denote the canonical HRF (here, a fixed double gamma function parameterized from previous studies). Then,

$$x_{ntq} = \int_0^{f_t} h(u) I_{nq}(f_t - u)\, du$$

where f_t is the time in seconds corresponding to the tth time point. This results in a time series corresponding to the assumed BOLD signal generated from the given HRF and stimulus pattern.

There is empirical support for the use of an autoregressive, or AR, model for the errors in (3.1) (Lindquist, 2008; Worsley et al., 2002). In a preliminary analysis of the ToM HCP data, we found that an AR(3) model was preferred for many locations, and

thus we will use an AR(3) model for all vertices. Let B denote the back-shift operator: $By_{nvt} = y_{n,v,t-1}$. Then the first-level model with time-series errors for the nth subject is

$$(1 - \phi_{nv1}B - \phi_{nv2}B^2 - \phi_{nv3}B^3)(y_{nvt} - \mathbf{x}'_{nt}\boldsymbol{e}_{nv} - \mathbf{z}'_{nt}\boldsymbol{\gamma}_{nv}) = \epsilon_{nvt}, \tag{3.3}$$

where $\epsilon_{nvt} \overset{iid}{\sim} \mathcal{N}(0, \tau^2_{nv})$. Here, τ^2_{nv} is the innovation variance whereas ξ^2_{nv} in (3.2) is the unconditional error variance.

3.2.2 Second level (population effects)

In the second-level, the subject-specific effects are generated from a fixed population effect plus a random effect. For each q, we have the linear model

$$e_{nvq} = \beta_{vq} + b_{nvq} \tag{3.4}$$

where $b_{nvq} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{b_q})$ for $n = 1, \ldots, N$ and $v = 1, \ldots, V$. We will also assume b_{nvq} and $b_{nvq'}$ are independent for all $q \neq q' \in \{1, \ldots, Q\}$.

Note that here we have assumed that the random effect variance, $\sigma^2_{b_1}, \ldots, \sigma^2_{b_Q}$, are constant across space, whereas the error variance, ξ^2_{nv}, varies across space. This contrasts with most presentations of the MUMM in which the index v is dropped because space is ignored, so implicitly, a separate variance component is estimated for each vertex (e.g., Worsley et al. 2002). However, it is consistent with extensions of the MUMM to spatial models in which a correlation structure of the random effects is estimated (e.g., Bowman 2007). This will be discussed further in Section 3.2.3.

The two-level model can be formalized in a single mixed model. We can substitute (3.4) into (3.3):

$$(1 - \phi_{nv1}B - \phi_{nv2}B^2 - \phi_{nv3}B^3)\left\{y_{nvt} - \mathbf{x}'_{nt}\left(\boldsymbol{\beta}_v + \boldsymbol{b}_{nv}\right) - \mathbf{z}'_{nt}\boldsymbol{\gamma}_{nv}\right\} = \epsilon_{nvt}$$

where $\boldsymbol{\beta}_v = [\beta_{v1}, \ldots, \beta_{vQ}]'$, $\boldsymbol{b}_{nv} = [b_{nv1}, \ldots, b_{nvQ}]'$ with $\boldsymbol{b}_{nv} \overset{iid}{\sim} N(\mathbf{0}, \mathbf{B})$ for $\mathbf{B} = \text{diag}(\sigma^2_{b_1}, \ldots, \sigma^2_{b_Q})$ and $n = 1, \ldots, N$, $v = 1, \ldots, V$. Additionally, we assume all b_{nvq} and ϵ_{nvt} are mutually independent.

Let $\mathbf{Y}_{nv} = [y_{nv1}, \ldots, y_{nvT}]'$ and $\mathbf{X}_n = [\mathbf{x}_{n1}, \ldots, \mathbf{x}_{nT}]' \in \mathbb{R}^{T \times Q}$. Then,

$$\text{Cov}\, \mathbf{Y}_{nv} = \mathbf{X}_{nt} \mathbf{B} \mathbf{X}'_{nt} + \xi^2_{nv} \boldsymbol{\Psi}_{nv}. \tag{3.5}$$

3.2.3 Estimating the MUMM

Arguably the most popular method for fitting the MUMM is using OLS to estimate subject-specific coefficients and then averaging these subject-specific coefficients to obtain estimators of the population parameters $\boldsymbol{\beta}_v$. This is called the summary statistics approach or the OLS approach. A survey of ninety fMRI papers found that 92% used the summary statistics approach (Mumford and Nichols, 2009). Alternatively, one could estimate all first and second-level parameters simultaneously using maximum likelihood (ML) or restricted maximum likelihood (REML). However, this is computationally very expensive when applied to hundreds of thousands of vertices in fMRI. Since the OLS estimators of $\boldsymbol{\beta}_v$ are statistically consistent, one can use coefficients from OLS but correct their standard errors by using the Yule-Walker equations to estimate the AR(p) parameters from the residuals. Worsley et al. (2002) observed that using residuals from OLS introduces bias into the estimates of the AR(p) parameters, and they propose a bias-reduction step that greatly improves estimation. They then construct an estimate of $\xi^2_{nv} \boldsymbol{\Psi}_{nv}$ and use generalized least squares to re-estimate the coefficients of the linear model. Finally, they adapt REML to estimate the random effect variance given the first-level parameters. Other approaches relax the assumption that the random-effect variance is constant across subjects. Then the GLS estimator of the fixed effects becomes a linear

combination of the subject-specific coefficients with weights determined by an estimate of the subject-specific random effect variance (Beckmann et al., 2003).

Perhaps surprisingly, the summary statistics estimators are nearly as powerful as estimators from a single-stage REML (Friston et al., 2005) or estimators from the weighted approach allowing for heterogeneous random effect variances (Mumford and Nichols, 2009) when one group of subjects is being analyzed. If there exist two groups with different random effect variances, then the weighted approach appears to improve estimation. Our data contains only one group of subjects, and the homogeneity assumption appears to be reasonable. Thus we use the summary statistics estimators described below. Their computational simplicity is very appealing. In fact, we don't even need to fit a time-series model in the first level, nor conduct temporal whitening. The summary statistics estimators are the default estimators when using Statistical Parametric Mapping (SPM) software.

Specifically, let $\mathbf{Z}_n = [\mathbf{z}_{n1}, \ldots, \mathbf{z}_{nT}]' \in \mathbb{R}^{T \times M}$ and let $\mathbf{X}_n^* = [\mathbf{X}_n, \mathbf{Z}_n]$. Define $\mathbf{K}_n^* = \mathbf{X}_n^* (\mathbf{X}_n^{*\prime} \mathbf{X}_n^*)^{-1}$. Then let $\mathbf{K}_n$ be the first Q columns of $\mathbf{K}_n^*$. Let $\hat{\boldsymbol{e}}_{nv}$ be an estimator of $\boldsymbol{e}_{nv} = [e_{nv1}, \ldots, e_{nvq}]'$. Throughout this manuscript, a "hat" over a random variable denotes an estimator of that random variable. Then

$$\hat{\boldsymbol{e}}_{nv} = \mathbf{K}_n' \mathbf{Y}_{nv}. \tag{3.6}$$

This notation is to keep track of the effect of non-orthogonal nuisance terms. A more straightforward exposition would assume the nuisance terms are orthogonal to the covariates of interest. Then (3.6) becomes $\hat{\boldsymbol{e}}_{nv} = (\mathbf{X}_n' \mathbf{X}_n)^{-1} \mathbf{X}_n' \mathbf{Y}_{nv}$. However, in applications they are not orthogonal.

Next, define the estimate of the population coefficients:

$$\hat{\boldsymbol{\beta}}_v = \frac{1}{N} \sum_{n=1}^{N} \hat{\boldsymbol{e}}_{nv}. \tag{3.7}$$

An estimator of the variance of (3.7) is also easy to calculate:

$$\widehat{\text{Cov}}\,\hat{\boldsymbol{\beta}}_v = \frac{1}{N(N-1)} \sum_{n=1}^{N} \left(\hat{\boldsymbol{e}}_{nv} - \hat{\boldsymbol{\beta}}_v\right)\left(\hat{\boldsymbol{e}}_{nv} - \hat{\boldsymbol{\beta}}_v\right)' \tag{3.8}$$

To gain insight into (3.7) and (3.8), we will examine the variance of (3.7) and the expected value of (3.8). First, note that

$$\text{Cov}\,\hat{\boldsymbol{\beta}}_v = \frac{1}{N^2} \sum_{n=1}^{N} \text{Cov}\,\hat{\boldsymbol{e}}_{nv}.$$

Next, note that

$$\begin{aligned}
&\text{E}\,\frac{1}{N(N-1)} \sum_{n=1}^{N} (\hat{\boldsymbol{e}}_{nv} - \hat{\boldsymbol{\beta}}_v)(\hat{\boldsymbol{e}}_{nv} - \hat{\boldsymbol{\beta}}_v)' = \\
&\frac{1}{N(N-1)} \sum_{n=1}^{N} \left\{\text{Cov}\,\hat{\boldsymbol{e}}_{nv} + \text{Cov}\,\hat{\boldsymbol{\beta}}_v - 2\text{Cov}\,(\hat{\boldsymbol{\beta}}_v, \hat{\boldsymbol{e}}_{nv})\right\} = \\
&\frac{1}{N(N-1)} \sum_{n=1}^{N} \left\{\text{Cov}\,\hat{\boldsymbol{e}}_{nv} + \frac{1}{N^2} \sum_{n'=1}^{N} \text{Cov}\,\hat{\boldsymbol{e}}_{n'v} - \frac{2}{N}\text{Cov}\,\hat{\boldsymbol{e}}_{nv}\right\} = \\
&\frac{1}{N(N-1)} \sum_{n=1}^{N} \frac{N-1}{N}\text{Cov}\,\hat{\boldsymbol{e}}_{nv} = \\
&\text{Cov}\,\hat{\boldsymbol{\beta}}_v.
\end{aligned}$$

Thus, the simple estimator in (3.8) is unbiased. Note that this estimation procedure never actually estimates the variance components of the model yet yields unbiased estimates of the variance of the fixed effect estimates.

One could consider the MUMM with vertex-specific random effect variances, $\mathbf{B}_v = \text{diag}(\sigma^2_{b_{v1}}, \ldots, \sigma^2_{b_{vQ}})$. Then one could use the same estimators. We note that previous spatial models have emphasized the increase in power relative to the MUMM. This may in part be due to the fact that a single variance component, $\mathbf{B}$, is estimated in the spatial models, whereas when the true model has a single random effect variance, the MUMM inefficiently estimates the variance of the fixed effects.

Some authors define the MUMM in terms of a statistical model in the first level, and then an equation for the second model where the estimators of the coefficients from the first level are the response variable (e.g., Penny et al. 2003). In their analysis of the summary statistics estimators, Beckmann et al. (2003) use this definition to attempt to explain the difference between the summary statistics "model" and a model allowing for subject-specific error variances. The results of Beckmann et al. (2003) can be summarized as follows. Suppose that the error variance is constant across subjects: $\xi^2_{nv} = \xi^2_{n'v}$ for all $n, n' \in 1, \ldots, N$. Then the summary statistics estimator of $\boldsymbol{\beta}_v$ is equivalent to the GLS estimator. Now let ξ^2_{nv} vary by subject. Then the GLS estimator under known covariance has lower variance than the summary statistics estimator. As previously noted, the improvements are generally minor for a study with a single group.

With respect to the MUMM, Woolrich et al. (2004) claim "there are no solutions in the frequentist literature to this model when the variance components are unknown." They suggest a hybrid approach wherein variance components are estimated using a Bayesian model with reference priors. We do not explore their method here, but note that their software is popular.

3.2.4 Applying t-tests for vertex-level inference

We are typically interested in whether vertices are differentially activated in one task versus another. Contrasts are used to test hypotheses of these types. Let $\hat{\boldsymbol{\beta}}_v \in \mathbb{R}^Q$, and let $\mathbf{c} = [c_1, \ldots, c_Q]$ such that $c_q = 1$ for the main effect of interest, $c_{q'} = -1$ for the contrast task, and $c_{q''} = 0$ otherwise. For example, in the HCP ToM analysis, we have $\hat{\boldsymbol{\beta}}_v \in \mathbb{R}^2$ with $c_1 = 1$ and $c_2 = -1$ corresponding to the contrast between the mentalizing

and random tasks (see Section 3.5.1 for additional information). Then a t-statistic is

$$t_v = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}}_v}{\sqrt{\mathbf{c}'\widehat{\mathrm{Cov}}\,(\hat{\boldsymbol{\beta}}_v)\mathbf{c}}}.$$

Under the null hypothesis that the contrast is equal to zero, this statistic is t-distributed with $N - 1$ degrees of freedom.

In this paper, we focus on the t-statistic from the main contrast. However, it is straightforward to define an F-statistic that simultaneously considers contrasts between the main effects of mentalizing and random, their time delay parameter derivatives, and their dispersal derivatives. This test is not signed, so it does not distinguish between areas that are significantly less activated by the task-of-interest and areas that are significantly more activated.

3.2.5 Region of Interest Mixed Model (ROIMM)

Now consider testing hypotheses regarding whether a group of vertices is activated. We can relax the assumption of spatial independence by averaging the BOLD signal across an ROI for each subject, and then conduct the usual mixed model analysis on the average signal. The ROI mixed model (ROIMM) replaces the flexibility in the MUMM with the assumption of homogeneity of main effects and AR(p) parameters within the ROI. Consider the rth region comprising vertices $\mathcal{V}_r = v_1, \ldots, v_{V_r}$. Let $\beta_q^{(r)}$ denote the overall activation in the rth region from the qth task. We will modify the MUMM to incorporate coefficients that are equivalent across the region. We will also introduce vertex-subject random effects. Consider an AR3 model,

$$(1 - \phi_{n1}B - \phi_{n2}B^2 - \phi_{n3}B^3)\left\{y_{nvt} - \mathbf{x}'_{nt}\left(\boldsymbol{\beta}^{(r)} + \boldsymbol{b}_{nv}\right) - \mathbf{z}'_{nt}\boldsymbol{\gamma}_{nv}\right\} = \epsilon_{nvt}$$

where $\boldsymbol{\beta}^{(r)} = [\beta_1^{(r)}, \ldots, \beta_Q^{(r)}]'$, and $\boldsymbol{b}_{nv} \overset{iid}{\sim} N\{\mathbf{0}, \text{diag}(\sigma^2_{b_1}, \ldots, \sigma^2_{b_Q})\}$ for $n = 1, \ldots, N$ and $v \in \mathcal{V}_r$. Now define $\boldsymbol{b}_n^q = [b_{n1q}, \ldots, b_{nVq}]'$ such that the ordering on the vertices is sequential for a given task. We specify a covariance structure for the random effects:

$$\boldsymbol{b}_n^q \sim \mathcal{N}(\mathbf{0}, \sigma^2_{b_q} \boldsymbol{\Omega}_q),$$

where $\boldsymbol{\Omega}_q$ is a $V_r \times V_r$ spatial correlation matrix. We will specify details of the covariance structure when we fit the spatiotemporal model, but for now, we maintain generality. Additionally, we assume all $\boldsymbol{b}_n^q$ and ϵ_{nvt} are mutually independent. Define

$$\mathbf{Y}_n^{(r)} = [\mathbf{Y}_{nv_1}, \ldots \mathbf{Y}_{nv_{V_r}}].$$

Then let

$$\bar{\mathbf{Y}}_n^{(r)} = \frac{1}{V_r} \left(\mathbf{1}'_{V_r} \otimes \mathbf{I}_T \right) \mathbf{Y}_n^{(r)}.$$

Let $\hat{\boldsymbol{e}}_n^{(r)} = \mathbf{K}_n' \bar{\mathbf{Y}}_n^{(r)}$, and define $\hat{\boldsymbol{\beta}}^{(r)}$ in a manner analogous to (3.7). Also define $\widehat{\text{Cov}}\, \hat{\boldsymbol{\beta}}^{(r)}$ in a manner analogous to (3.8).

3.3 A Spatiotemporal Mixed Effects Model (STMM)

3.3.1 Model formulation

We will assume there exists some parcellation that defines independent regions. Then we can treat each region as a separate estimation problem, which makes model estimation computationally feasible. In this framework, information from one region does not provide any information on another region. Ultimately, we will create a full covariance matrix for all vertices, where the covariance between vertices in different regions is assumed to be zero. Then we can conduct ROI inference over any set of vertices (including vertices in different regions).

To simplify the notation, let us consider a single region and drop the superscript r. Let

$$\mathbf{Y}_n = [y_{n11}, \ldots, y_{n1T}, y_{n21}, \ldots, y_{n2T}, \ldots, y_{nVT}]'$$

denote the BOLD signal for the nth subject at the vertices within a region, where with slight abuse of notation, V is the number of vertices in the region. Let $\mathbf{Y} = [\mathbf{Y}_1', \ldots, \mathbf{Y}_N']'$.

Recall that $\mathbf{X}_n \in \mathbb{R}^{T \times Q}$ (covariates for which we wish to conduct population-level inference) and $\mathbf{Z}_n \in \mathbb{R}^{T \times M}$ (nuisance covariates). Let $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_Q]$ denote a vector of fixed effects, which are constant across subjects and across space. Let $\boldsymbol{\gamma}_{nv} = [\gamma_{nv1}, \ldots, \gamma_{nvM}]'$ denote the fixed effects associated with the nuisance terms for the nth subject at the vth vertex. Let $\boldsymbol{s}_n = [s_{n1}, \ldots, s_{nQ}]'$ denote the subject-specific random slopes and $\boldsymbol{s} = [\boldsymbol{s}_1', \ldots, \boldsymbol{s}_N']'$. Note that the subject random effect captures a baseline correlation in the BOLD signal between vertices within the same region in the same subject (since we define a different model for each region). Next define the population-level vertex-wise random effect: $\boldsymbol{u}_v = [u_{v1}, \ldots, u_{vQ}]'$ and $\boldsymbol{u} = [\boldsymbol{u}_1', \ldots, \boldsymbol{u}_V']'$. Define $\boldsymbol{u}^q = [u_{1q}, \ldots, u_{Vq}]'$ such that the ordering of the vertices is sequential. Define the interaction between the vertex and subject effects: $\boldsymbol{b}_{nv} = [b_{nv1}, \ldots, b_{nvQ}]'$; $\boldsymbol{b}_n = [\boldsymbol{b}_{n1}', \ldots, \boldsymbol{b}_{nV}']'$; $\boldsymbol{b} = [\boldsymbol{b}_1', \ldots, \boldsymbol{b}_V']'$; and $\boldsymbol{b}_n^q = [b_{n1q}, \ldots, b_{nVq}]'$. Define the errors $\boldsymbol{a}_{nv} = [a_{nv1}, \ldots, a_{nvT}]'$; $\boldsymbol{a}_n = [\boldsymbol{a}_{n1}', \ldots, \boldsymbol{a}_{nV}']'$; and $\boldsymbol{a} = [\boldsymbol{a}_1', \ldots, \boldsymbol{a}_N']'$.

Following the notation defined on p.445 in Searle et al. (2009), for some arbitrary matrices $\mathbf{A}_1, \ldots, \mathbf{A}_N$, let

$$\{_c\ \mathbf{A}_n\}_{n=1}^N = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_N \end{bmatrix},$$

where the c denotes that we are stacking matrices column-wise. We will also use $\{_r\ \mathbf{A}_n\}_{n=1}^N$ to denote row-wise concatenation. Let $\oplus_{n=1}^N \mathbf{A}_n$ denote the direct sum, i.e.,

$\text{diag}\{\mathbf{A}_1, \ldots, \mathbf{A}_N\}$. (See also Appendix C.1 for a summary of matrix notation.)

We define a spatiotemporal mixed model (STMM):

$$y_{nvt} = \mathbf{x}'_{nt}(\boldsymbol{\beta} + \boldsymbol{u}_v + \boldsymbol{s}_n + \boldsymbol{b}_{nv}) + \mathbf{z}'_{nt}\boldsymbol{\gamma}_{nv} + a_{nvt}, \tag{3.9}$$

which in matrix form is

$$\begin{aligned}\mathbf{Y} = \{_c\ \mathbf{1}_V \otimes \mathbf{X}_n\}_{n=1}^N\ \boldsymbol{\beta} + \{_c\ \mathbf{I}_V \otimes \mathbf{X}_n\}_{n=1}^N\ \boldsymbol{u} + \left[\oplus_{n=1}^N (\mathbf{1}_V \otimes \mathbf{X}_n)\right] \boldsymbol{s} \\ + \left[\oplus_{n=1}^N (\mathbf{I}_V \otimes \mathbf{X}_n)\right] \boldsymbol{b} + \left[\oplus_{n=1}^N (\mathbf{I}_V \otimes \mathbf{Z}_n)\right] \boldsymbol{\gamma} + \boldsymbol{a},\end{aligned}$$

where

$$\boldsymbol{u}^q \sim \mathcal{N}(\mathbf{0}, \sigma^2_{u_q}\boldsymbol{\Gamma}_q);$$

$$\boldsymbol{s}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{S})$$

with $\mathbf{S} = \text{diag}(\sigma^2_{s_1}, \ldots, \sigma^2_{s_Q})$;

$$\boldsymbol{b}^q_n \sim \mathcal{N}(\mathbf{0}, \sigma^2_{b_q}\boldsymbol{\Omega}_q);$$

and

$$\boldsymbol{a} \sim \mathcal{N}(\mathbf{0}, \oplus_{n=1}^N \oplus_{v=1}^V \xi^2_{nv}\boldsymbol{\Psi}_{nv}).$$

The spatial correlation matrices, $\boldsymbol{\Gamma}_q$ and $\boldsymbol{\Omega}_q$ of $\boldsymbol{u}_q$ and $\boldsymbol{b}_q$, respectively, are discussed below. As in the MUMM and (3.2), $\boldsymbol{\Psi}_{nv}$ is the autocorrelation matrix for an AR(3) process. Note we have assumed that the variances of the spatial random effects are constant across a region while spatial differences in variance are captured by the error variance, $\xi^2_{nv}\boldsymbol{\Psi}_{nv}$, $v = 1, \ldots, V$, $n = 1, \ldots, N$. Additionally, we assume all $\boldsymbol{u}^q$, $\boldsymbol{b}^q_n$, $\boldsymbol{s}_n$, and $\boldsymbol{a}_n$ for $n = 1, \ldots, N$ and $q = 1, \ldots, Q$ are mutually independent. This model implies a non-separable covariance matrix because the error variances vary across space.

The model can be formulated as a hierarchical model. In the first level,

$$y_{nvt} = \mathbf{x}'_{nt}\boldsymbol{e}_{nv} + \mathbf{z}'_{nt}\boldsymbol{\gamma}_{nv} + a_{nvt};$$

then

$$e_{nvq} = \beta_q + u_{vq} + s_{nq} + b_{nvq},\ q = 1, \ldots, Q. \tag{3.10}$$

Let $\|v_i - v_j\|$ denote a distance between vertices v_i and v_j and $\mathbf{\Gamma}_{q;v_i,v_j}$ denote the corresponding element of $\mathbf{\Gamma}_q$. We will additionally assume that $\mathbf{\Gamma}_{q;v_i,v_j} = \mathbf{\Gamma}_{q;v_k,v_l}$ for $\|v_i - v_j\| = \|v_k - v_l\|$. When defining our estimators of spatial correlation, we will also assume that $\mathbf{\Omega}_q = \mathbf{\Gamma}_q$, although we will maintain generality in this section.

Bowman (2007) examined a variety of variograms for volume-based fMRI data and found that the exponential variogram was most supported. Our exploration of the HCP data on the cortical surface for the Gordon networks also indicate this model is appropriate. The exponential covariogram is defined

$$\operatorname{Cov} b_v, b_{v'} = \begin{cases} l_0 + \lambda_1 & v = v' \\ \lambda_1 e^{-\theta\|v - v'\|} & v \neq v'. \end{cases}$$

where l_0 is the nugget effect, which is equal to the micro-scale variance plus the variance due to measurement error. Our hierarchical model in fact includes a spatially varying micro-scale and measurement error variance component via ξ^2_{nv}, so we assume the nugget effect for the random effects is equal to zero. We have

$$\mathbf{\Gamma}_{q;v,v'} = e^{-\theta_{u_q}\|v - v'\|}$$

and

$$\mathbf{\Omega}_{q;v,v'} = e^{-\theta_{b_q}\|v - v'\|}.$$

Let $\psi^{(nv)}_{t,t'}$ be equal to the corresponding element of $\mathbf{\Psi}_{nv}$. Let $\mathbf{U} = \operatorname{diag}(\sigma^2_{u_1}, \ldots, \sigma^2_{u_Q})$ and $\mathbf{B} = \operatorname{diag}(\sigma^2_{b_1}, \ldots, \sigma^2_{b_Q})$. Let $\mathbf{\Gamma}_{v,v'} = \operatorname{diag}(\mathbf{\Gamma}_{1;v,v'}, \ldots, \mathbf{\Gamma}_{Q;v,v'})$. The model implies

$$\operatorname{Cov} y_{nvt}, y_{n',v',t'} = \begin{cases} \mathbf{x}'_{nt}(\mathbf{U} + \mathbf{S} + \mathbf{B})\mathbf{x}_{nt'} + \xi^2_{nv}\psi^{(nv)}_{t,t'} & n = n';\, v = v';\ \text{any } t, t' \\ \mathbf{x}'_{nt}(\mathbf{U}\mathbf{\Gamma}_{v,v'} + \mathbf{S} + \mathbf{B}\mathbf{\Omega}_{v,v'})\mathbf{x}_{nt'} & n = n';\, v \neq v';\ \text{ any } t, t' \\ \mathbf{x}'_{nt}\mathbf{U}\mathbf{\Gamma}_{v,v'}\mathbf{x}_{n't'} & n \neq n';\ \text{any } v, v';\ \text{ any } t, t' \end{cases}$$

Note that none of the observations are independent – even observations from different subjects are correlated via $\boldsymbol{u}$.

For computational reasons, we will project the full data matrix $\mathbf{Y}$ onto the space spanned by the covariates. Recall from (3.6) that $\mathbf{X}_n^* = [\mathbf{X}_n, \mathbf{Z}_n]$, $\mathbf{K}_n^* = \mathbf{X}_n^*(\mathbf{X}_n^{*\prime}\mathbf{X}_n^*)^{-1}$, and $\mathbf{K}_n$ comprises the first Q columns of $\mathbf{K}_n^*$. Then define

$$\boldsymbol{d}_{nv} = \mathbf{K}_n{}'\mathbf{Y}_{nv}. \tag{3.11}$$

Let $\boldsymbol{d}_n = [\boldsymbol{d}_{n1}', \ldots, \boldsymbol{d}_{nV}']'$; let $\boldsymbol{d} = [\boldsymbol{d}_1', \ldots, \boldsymbol{d}_N']'$; and let $\boldsymbol{k}_{nv} = \mathbf{K}_n'\boldsymbol{a}_{nv}$. Then we have the second-level model

$$\boldsymbol{d} = (1_N \otimes 1_V \otimes \mathbf{I}_Q)\boldsymbol{\beta} + (1_N \otimes \mathbf{I}_V \otimes \mathbf{I}_Q)\boldsymbol{u} + (\mathbf{I}_N \otimes 1_V \otimes \mathbf{I}_Q)\boldsymbol{s} + (\mathbf{I}_N \otimes \mathbf{I}_V \otimes \mathbf{I}_Q)\boldsymbol{b} + \{_c\ \{_c\ \boldsymbol{k}_{nv}\}_{v=1}^V\}_{n=1}^N.$$

Note that $\boldsymbol{d}_{nv}$ is equivalent to the output from a first-level analysis using OLS estimators. However, it serves a conceptually different role here. In the MUMM, $\hat{\boldsymbol{e}}_{nv}$ in (3.6) was an estimator of $\boldsymbol{e}_{nv}$; here, we are treating $\boldsymbol{d}_{nv}$ as a dimension-reducing transformation of $\mathbf{Y}_{nv}$, and we are keeping track of the transformed error $\boldsymbol{a}_{nv}$. Put another way, $\boldsymbol{e}_{nv}$ in the STMM was defined in (3.10): $e_{nvq} = \beta_{vq} + u_{vq} + s_{nq} + b_{nvq}$, which is different from $d_{nvq} = \beta_v + u_{vq} + s_{nq} + b_{nvq} + k_{nvq}$ where k_{nvq} is the qth element of the transformed error, $\boldsymbol{k}_{nv}$. The choice of transformation is addressed in the discussion section.

Let $\boldsymbol{\Sigma} \in \mathbb{R}^{NVQ \times NVQ}$ denote the covariance matrix of $\boldsymbol{d}$. Let $\boldsymbol{\Gamma}$ be the $VQ \times VQ$ *covariance* matrix comprising $\oplus_{q=1}^Q \sigma_{u_q}^2 \boldsymbol{\Gamma}_q$ permuted to correspond to the indexing in $\boldsymbol{d}_n$; and similarly define $\boldsymbol{\Omega}$. Let $\mathbf{J}_N = 1_N 1_N'$. Then,

$$\boldsymbol{\Sigma} = \mathbf{J}_N \otimes \boldsymbol{\Gamma} + \mathbf{I}_N \otimes \mathbf{J}_V \otimes \mathbf{S} + \mathbf{I}_N \otimes \boldsymbol{\Omega} + \oplus_{n=1}^N \oplus_{v=1}^V \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n. \tag{3.12}$$

3.3.2 Estimating the variance components

Estimating the covariance of the errors

First, we estimate ξ_{nv}^2 and $\mathbf{\Psi}_{nv}$ using the first-level residuals. Let $\mathbf{H}_n = \mathbf{X}_n^* \mathbf{K}_n^{*\prime}$ (i.e., the hat matrix) and let $\hat{\boldsymbol{r}}_{nv} = (\mathbf{I}_T - \mathbf{H}_n)\mathbf{Y}_{nv}$. Note that

$$\begin{aligned}\operatorname{Cov} \hat{\boldsymbol{r}}_{nv} &= (\mathbf{I}_T - \mathbf{H}_n)\left\{\mathbf{X}_n (\mathbf{U} + \mathbf{S} + \mathbf{B}\} \mathbf{X}_n' + \xi_{nv}^2 \mathbf{\Psi}_{nv}\right)(\mathbf{I}_T - \mathbf{H}_n) \\ &= \xi_{nv}^2 (\mathbf{I}_T - \mathbf{H}_n) \mathbf{\Psi}_{nv} (\mathbf{I}_T - \mathbf{H}_n) .\end{aligned}$$

Our goal is to estimate $\xi_{nv}^2 \mathbf{\Psi}_{nv}$. One approach would be to use $\hat{\boldsymbol{r}}_{nv}$ to estimate the sample autocorrelations, use the Yule-Walker equations to estimate the innovation variance and the AR coefficients, and then calculate the unconditional variance and correlation matrix. As noted in Worsley et al. (2002), this approach produces biased estimates.

It can be shown that the bias of the OLS estimator of the error variance increases as positive dependence increases. We have

$$\frac{1}{T - (Q+M)} \mathrm{E} \sum_{t=1}^{T} (y_{nvt} - \hat{y}_{nvt})^2 = \frac{1}{T - (Q+M)} \xi_{nv}^2 \{T - \operatorname{tr}(\mathbf{\Psi}_{nv} \mathbf{H}_n)\} . \tag{3.13}$$

See Appendix C.3 for details. Note $\operatorname{tr}(\mathbf{H}_n) = (Q + M)$, so the OLS estimator is unbiased for iid errors. Positive off-diagonal elements of $\mathbf{\Psi}_{nv}$ decrease the expected value of the OLS estimator, resulting in downward bias.

First, we derive reduced-biased estimates of the sample autocorrelation function. For $\ell \in \{0, \ldots, T-1\}$, define an upper triangular $T \times T$ matrix $\mathbf{D}_\ell$ such that $(\mathbf{D}_\ell)_{ij} = 1$ for $j = i + \ell$ with $i = \ell + 1, \ldots, T - \ell$, and zero elsewhere. Then consider the cross-product between $\hat{r}_{nvt}$ and $\hat{r}_{nv,t-\ell}$:

$$\sum_{t=\ell+1}^{T} \hat{r}_{nvt}, \hat{r}_{nv,t-\ell} = \hat{\boldsymbol{r}}_{nv}' \mathbf{D}_\ell \hat{\boldsymbol{r}}_{nv}.$$

Then

$$\begin{aligned}
\mathrm{E}\,\hat{\boldsymbol{r}}'_{nv}\mathbf{D}_\ell\hat{\boldsymbol{r}}_{nv} &= \mathrm{E}\,\mathrm{tr}(\hat{\boldsymbol{r}}'_{nv}\mathbf{D}_\ell\hat{\boldsymbol{r}}_{nv}) \\
&= \mathrm{tr}(\mathbf{D}_\ell \mathrm{E}\,\hat{\boldsymbol{r}}_{nv}\hat{\boldsymbol{r}}'_{nv}) \\
&= \mathrm{tr}(\mathbf{D}_\ell \mathrm{Cov}\,\hat{\boldsymbol{r}}_{nv}) \\
&= \mathrm{tr}\Big\{\mathbf{D}_\ell(\mathbf{I}_T - \mathbf{H}_n)\boldsymbol{\Psi}_{nv}(\mathbf{I}_T - \mathbf{H}_n)\xi^2_{nv}\Big\} \\
&= \mathrm{tr}\Big\{(\mathbf{I}_T - \mathbf{H}_n)\mathbf{D}_\ell(\mathbf{I}_T - \mathbf{H}_n)\boldsymbol{\Psi}_{nv}\xi^2_{nv}\Big\}.
\end{aligned}$$

Now let $\psi_{nv}(\ell)$ denote the autocorrelation at lag ℓ, and let $\rho_{nv}(\ell) = \xi^2_{nv}\psi_{nv}(\ell)$. Note that

$$\boldsymbol{\Psi}_{nv}\xi^2_{nv} = \rho_{nv}(0)\mathbf{I}_T + \sum_{t=1}^{T-1}\rho_{nv}(t)(\mathbf{D}_t + \mathbf{D}'_t).$$

Then we have

$$\mathrm{E}\,\hat{\boldsymbol{r}}'_{nv}\mathbf{D}_\ell\hat{\boldsymbol{r}}_{nv} = \mathrm{tr}\left[(\mathbf{I}_T - \mathbf{H}_n)\mathbf{D}_\ell(\mathbf{I}_T - \mathbf{H}_n)\left\{\rho_{nv}(0)\mathbf{I}_T + \sum_{t=1}^{T-1}\rho_{nv}(t)(\mathbf{D}_t + \mathbf{D}'_t)\right\}\right].$$

From the above formula, we define a system of T equations and T unknowns for $\rho_{nv}(\ell)$, $\ell = 0, \ldots, T-1$. By replacing the left-hand side with $\hat{\boldsymbol{r}}'_{nv}\mathbf{D}_\ell\hat{\boldsymbol{r}}_{nv}$, we can solve for $\rho_{nv}(0), \ldots, \rho_{nv}(T-1)$. In practice, we will only derive estimates of the first L autocovariances, where L is a number chosen such that there is an adequate number of terms in $\hat{\boldsymbol{r}}'_{nv}\mathbf{D}_\ell\hat{\boldsymbol{r}}_{nv}$. Then define the $(L+1)\times(L+1)$ matrix $\mathbf{M}_n$, where for $\ell = 0, \ldots, L$,

$$[\mathbf{M}_n]_{\ell+1,j+1} = \begin{cases} \mathrm{tr}\{(\mathbf{I}_T - \mathbf{H}_n)\mathbf{D}_\ell\} & \text{if } j = 0, \\ \mathrm{tr}\Big\{(\mathbf{I}_T - \mathbf{H}_n)\mathbf{D}_\ell(\mathbf{I}_T - \mathbf{H}_n)(\mathbf{D}_j + \mathbf{D}'_j)\Big\} & \text{if } j = 1, \ldots, L. \end{cases} \tag{3.14}$$

Let $\widehat{\mathbf{R}}_{nv} = [\hat{\boldsymbol{r}}'_{nv}\mathbf{D}_0\hat{\boldsymbol{r}}_{nv}, \ldots, \hat{\boldsymbol{r}}'_{nv}\mathbf{D}_L\hat{\boldsymbol{r}}_{nv}]'$. Let $\hat{\boldsymbol{\rho}}_{nv} = [\hat{\rho}_{nv}(0), \ldots, \hat{\rho}_{nv}(L)]'$. Then we have

$$\hat{\boldsymbol{\rho}}_{nv} = \mathbf{M}_n^{-1}\widehat{\mathbf{R}}_{nv}. \tag{3.15}$$

Note that we only need to calculate $\mathbf{M}_n^{-1}$ once (for each subject).

Our estimate of the unconditional variance is then $\hat{\xi}^2_{nv} = \hat{\rho}_{nv}(0)$, and we obtain reduced-biased autocorrelation estimates $\hat{\psi}_{nv}(\ell) = \hat{\rho}_{nv}(\ell)/\hat{\rho}_{nv}(0)$. We then use the Yule-

Walker equations to obtain $\hat{\phi}_{nv1}, \ldots, \hat{\phi}_{nv3}$. Then $\hat{\psi}_{nv}(1), \ldots, \hat{\psi}_{nv}(3)$, and $\hat{\phi}_{nv1}, \ldots, \hat{\phi}_{nv3}$ are used to construct the $T \times T$ autocorrelation matrix (Shumway and Stoffer, 2010).

The choice of L does not appear to greatly affect estimation. Worsley et al. (2002) choose $L = p$, where p is the AR order. We found that $L = 20$ works well in practice. Increases beyond L were not found to reduce the bias in simulations.

For the sections that follow, it is convenient to define

$$MSR = \frac{1}{NV} \sum_{n=1}^{N} \sum_{v=1}^{V} \hat{\xi}_{nv}^2 \mathbf{K}_n' \widehat{\mathbf{\Psi}}_{nv} \mathbf{K}_n,$$

which will serve a role similar to the mean square residual in ANOVA decompositions.

Estimating the variance of the subject-vertex interaction effect

We next turn to estimating the remaining variance components. Define the following quantities:

$$\bar{\boldsymbol{d}}_{\cdot v} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{d}_{nv}$$
$$\bar{\boldsymbol{d}}_{n\cdot} = \frac{1}{V} \sum_{v=1}^{V} \boldsymbol{d}_{nv}$$
$$\bar{\boldsymbol{d}}_{\cdot\cdot} = \frac{1}{NV} \sum_{n=1}^{N} \sum_{v=1}^{V} \boldsymbol{d}_{nv}.$$

Let

$$g_q = \sum_{v=1}^{V} \sum_{v'=1}^{V} \mathbf{\Gamma}_{q;v,v'}$$

and $\mathbf{G} = \text{diag}(g_1, \ldots, g_Q)$. Then let

$$w_q = \sum_{v=1}^{V} \sum_{v'=1}^{V} \mathbf{\Omega}_{q;v,v'}$$

and $\mathbf{W} = \text{diag}(w_1, \ldots, w_Q)$.

Towards estimating $\mathbf{B}$, consider a measure of the mean square due to the interaction between subject and vertex:

$$MSB = \frac{1}{(N-1)(V-1)} \sum_{n=1}^{N} \sum_{v=1}^{V} (\boldsymbol{d}_{nv} - \bar{\boldsymbol{d}}_{n\cdot} - \bar{\boldsymbol{d}}_{\cdot v} + \bar{\boldsymbol{d}}_{\cdot\cdot})(\boldsymbol{d}_{nv} - \bar{\boldsymbol{d}}_{n\cdot} - \bar{\boldsymbol{d}}_{\cdot v} + \bar{\boldsymbol{d}}_{\cdot\cdot})'.$$

Here, MSB is a $Q \times Q$ matrix. In our application, our tasks are nearly orthogonal, and we assume the random effects associated with each task are independent. Then we only calculate the diagonal elements.

Let us first consider the case where $q = 1$. Let $\bar{\mathbf{J}}_N = \frac{1}{N}\mathbf{J}$. We can define the matrix $\mathbf{C}_N \otimes \mathbf{C}_V$ where $\mathbf{C}_N = \mathbf{I}_N - \bar{\mathbf{J}}_N$. Then

$$MSB = \frac{1}{(N-1)(V-1)} \boldsymbol{d}'(\mathbf{C}_N \otimes \mathbf{C}_V)\boldsymbol{d}$$

When $q = 1$, the computation of $\mathrm{E}\, MSB$ simplifies considerably using the fact that

$$\mathrm{E}\, MSB = \frac{1}{(N-1)(V-1)} \left[\mathrm{tr}\{(\mathbf{C}_N \otimes \mathbf{C}_V)\boldsymbol{\Sigma}\} + \boldsymbol{\mu}'(\mathbf{C}_N \otimes \mathbf{C}_V)\boldsymbol{\mu}\right]$$

where $\boldsymbol{\mu} = \beta \mathbf{1}_{NV}$. The second term is equal to zero. Then recalling the covariance matrix in (3.12),

$$\begin{aligned}
&\mathrm{tr}\Big\{(\mathbf{C}_N \otimes \mathbf{C}_V)(\mathbf{J}_N \otimes \sigma_u^2 \boldsymbol{\Gamma}_1 + \mathbf{I}_N \otimes \sigma_s^2 \mathbf{J}_V + \mathbf{I}_N \otimes \sigma_b^2 \boldsymbol{\Omega}_1 + \oplus_{n=1}^{N} \oplus_{v=1}^{V} \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n)\Big\} = \\
&\mathrm{tr}\, \mathbf{0} + \mathrm{tr}\, \mathbf{0} + \sigma_b^2 \mathrm{tr}(\mathbf{C}_N)\mathrm{tr}(\mathbf{C}_V \boldsymbol{\Omega}_1) + \\
&\mathrm{tr}\Big\{\Big(\mathbf{I}_{NV} - \bar{\mathbf{J}}_N \otimes \mathbf{I}_V - \mathbf{I}_N \otimes \bar{\mathbf{J}}_V + \bar{\mathbf{J}}_{NV}\Big) \oplus_{n=1}^{N} \oplus_{v=1}^{V} \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n)\Big\} = \\
&\sigma_b^2 (N-1)\left(V - \frac{w_1}{V}\right) + \left(1 - \frac{1}{N} - \frac{1}{V} + \frac{1}{NV}\right) \sum_{n=1}^{N} \sum_{v=1}^{V} \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n,
\end{aligned}$$

and after a little more algebra,

$$\mathrm{E}\, MSB = \sigma_b^2 \left(\frac{V}{V-1} - \frac{w_1}{V(V-1)}\right) + \frac{1}{NV} \sum_{n=1}^{N} \sum_{v=1}^{V} \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n.$$

More generally, we can derive estimators for a vector response $\boldsymbol{d}_{nv} \in \mathbb{R}^Q$, corresponding to multiple tasks. Details are provided in Appendix C.4.

$$\mathrm{E}\, MSB = \left(\frac{V}{V-1}\mathbf{I}_Q - \frac{1}{V(V-1)}\mathbf{W}\right)\mathbf{B} + \frac{1}{NV}\sum_{n=1}^{N}\sum_{v=1}^{V}\xi_{nv}^2\mathbf{K}_n'\boldsymbol{\Psi}_{nv}\mathbf{K}_n. \tag{3.16}$$

Note that under zero spatial and temporal dependence with $\mathbf{Z}_n = 0$, $\mathbf{X}_n = \mathbf{1}_T$ for all n, and $\xi_{nv}^2 = \sigma_a^2$ for all n, v, we have $\mathrm{E}\, MSB = \sigma_b^2 + \frac{1}{T}{\sigma_a}^2$. Then the model resembles the classic two-way crossed random effects model (e.g., p.123, Searle et al. 2009).

We define the estimator

$$\widehat{\mathbf{B}} = \left(\frac{V}{V-1}\mathbf{I}_Q - \frac{1}{V(V-1)}\widehat{\mathbf{W}}\right)^{-1}(MSB - MSR).$$

As is the case for method of moments estimators in other mixed models, $\hat{\sigma}_{b_q}^2$ can be negative, in which case we replace it with $1e-06$ (effectively zero for the scaling in our simulations and data application). This introduces some bias into our estimator but decreases its mean squared error (p.130, Searle et al. 2009).

Estimating the variance of the subject random effect

Towards estimating $\mathbf{S}$, consider the mean square due to subject:

$$MSS = \frac{1}{N-1}\sum_{n=1}^{N}\sum_{v=1}^{V}(\bar{\boldsymbol{d}}_{n\cdot} - \bar{\boldsymbol{d}}_{\cdot\cdot})(\bar{\boldsymbol{d}}_{n\cdot} - \bar{\boldsymbol{d}}_{\cdot\cdot})'.$$

Note that

$$\mathrm{E}\,(\bar{\boldsymbol{d}}_{n\cdot} - \bar{\boldsymbol{d}}_{\cdot\cdot})(\bar{\boldsymbol{d}}_{n\cdot} - \bar{\boldsymbol{d}}_{\cdot\cdot})' = \mathrm{Cov}\,\bar{\boldsymbol{d}}_{n\cdot} - 2\mathrm{Cov}\,(\bar{\boldsymbol{d}}_{n\cdot}, \bar{\boldsymbol{d}}_{\cdot\cdot}) + \mathrm{Cov}\,\bar{\boldsymbol{d}}_{\cdot\cdot}.$$

Using the results in (C.3), we have

$$
\begin{aligned}
&\mathrm{E}\,(\bar{\boldsymbol{d}}_{n\cdot} - \bar{\boldsymbol{d}}_{\cdot\cdot})(\bar{\boldsymbol{d}}_{n\cdot} - \bar{\boldsymbol{d}}_{\cdot\cdot})' \\
&= \mathbf{S} + \frac{1}{V^2}\mathbf{GU} + \frac{1}{V^2}\mathbf{WB} + \frac{1}{V^2}\sum_{v=1}^{V} \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n \\
&\quad - 2\left(\frac{1}{N}\mathbf{S} + \frac{1}{V^2}\mathbf{GU} + \frac{1}{NV^2}\mathbf{WB} + \frac{1}{NV^2}\sum_{v=1}^{V} \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n\right) \\
&\quad + \frac{1}{N}\mathbf{S} + \frac{1}{V^2}\mathbf{GU} + \frac{1}{NV^2}\mathbf{WB} + \frac{1}{N^2V^2}\sum_{n=1}^{N}\sum_{v=1}^{V} \mathbf{K}_n \boldsymbol{\Psi}_{nv} \mathbf{K}_n' \\
&= \frac{N-1}{N}\mathbf{S} + \frac{N-1}{NV^2}\mathbf{WB} + \frac{N-2}{NV^2}\sum_{v=1}^{V} \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n + \frac{1}{N^2V^2}\sum_{n=1}^{N}\sum_{v=1}^{V} \mathbf{K}_{n'}' \boldsymbol{\Psi}_{nv} \mathbf{K}_{n'}.
\end{aligned}
$$

Putting this together, we have

$$
\mathrm{E}\,MSS = V\mathbf{S} + \frac{1}{V}\mathbf{WB} + \frac{1}{NV}\sum_{n=1}^{N}\sum_{v=1}^{V} \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n.
$$

In the univariate case under no spatial or temporal dependence, the result parallels the expected mean square of a random effects model for a two-factorial crossed design. Namely, if $\boldsymbol{\Omega} = \mathbf{I}_V$ and $\xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n = \frac{1}{T}\sigma_a^2$, then $\mathrm{E}\,MSS = V\sigma_s^2 + \sigma_b^2 + \frac{1}{T}\sigma_a^2$.

We propose the estimator

$$
\widehat{\mathbf{S}} = \frac{1}{V}MSS - \frac{1}{V^2}\widehat{\mathbf{W}}\widehat{\mathbf{B}} - \frac{1}{V}MSR. \tag{3.17}
$$

As in the case of $\widehat{\mathbf{B}}$, if some $\hat{\sigma}_{s_q}^2 < 0$, then we replace it with $1e-06$. From (3.17), it can be seen that an increase in (positive) spatial dependence results in a decrease in the subject-specific variance component.

Estimating the variance of the vertex random effect

Next define

$$
MSU = \frac{1}{V-1}\sum_{n=1}^{N}\sum_{v=1}^{V} (\bar{\boldsymbol{d}}_{\cdot v} - \bar{\boldsymbol{d}}_{\cdot\cdot})(\bar{\boldsymbol{d}}_{\cdot v} - \bar{\boldsymbol{d}}_{\cdot\cdot})'.
$$

Then,

$$\sum_{v=1}^{V} \mathrm{E}\,(\bar{\boldsymbol{d}}_{\cdot v} - \bar{\boldsymbol{d}}_{\cdot\cdot})(\bar{\boldsymbol{d}}_{\cdot v} - \bar{\boldsymbol{d}}_{\cdot\cdot})' =$$

$$V\mathbf{U} - \frac{1}{V}\mathbf{G}\mathbf{U} + \frac{V}{N}\mathbf{B} - \frac{1}{NV}\mathbf{W}\mathbf{B} + \frac{1}{N^2}\sum_{n=1}^{N}\sum_{v=1}^{V}\xi_{nv}^2\mathbf{K}_n'\boldsymbol{\Psi}_{nv}\mathbf{K}_n - \frac{1}{N^2V}\sum_{n=1}^{N}\sum_{v'=1}^{V}\mathbf{K}_n\boldsymbol{\Psi}_{nv'}\mathbf{K}_n'$$

$$= \left(V\mathbf{I}_Q - \frac{1}{V}\mathbf{G}\right)\mathbf{U} + \left(\frac{V}{N}\mathbf{I}_Q - \frac{1}{NV}\mathbf{W}\right)\mathbf{B} + \frac{V-1}{N^2V}\sum_{n=1}^{N}\sum_{v'=1}^{V}\mathbf{K}_n\boldsymbol{\Psi}_{nv'}\mathbf{K}_n'.$$

and we have

$$\begin{aligned}\mathrm{E}\,MS\,U &= \left(\frac{NV}{V-1}\mathbf{I}_Q - \frac{N}{V(V-1)}\mathbf{G}\right)\mathbf{U} + \Bigg(\frac{V}{V-1}\mathbf{I}_Q \\ &\quad - \frac{1}{V(V-1)}\mathbf{W}\Bigg)\mathbf{B} + \frac{1}{NV}\sum_{n=1}^{N}\sum_{v'=1}^{V}\mathbf{K}_n\boldsymbol{\Psi}_{nv'}\mathbf{K}_n' \\ &= \left(\frac{NV}{V-1}\mathbf{I}_Q - \frac{N}{V(V-1)}\mathbf{G}\right)\mathbf{U} + \mathrm{E}\,MS\,B.\end{aligned} \tag{3.18}$$

Note in the univariate case with no spatial or temporal dependence and $\xi_{nv}^2\mathbf{K}_n'\boldsymbol{\Psi}_{nv}\mathbf{K}_n = \frac{1}{T}\sigma_a^2$, then $\mathrm{E}\,MS\,U = N\sigma_u^2 + \sigma_b^2 + \frac{1}{T}\sigma_a^2$.

Recalling (3.16), we define the estimator:

$$\widehat{\mathbf{U}} = \left(\frac{NV}{V-1}\mathbf{I}_Q - \frac{N}{V(V-1)}\mathbf{G}\right)^{-1}(MS\,U - MS\,B). \tag{3.19}$$

As with the previous variance component estimators, if $\hat{\sigma}_{u_q}^2 < 0$, we replace it with $1e-06$.

3.3.3 Estimating spatial dependence parameters

A popular tool for assessing spatial dependence is the empirical variogram. Let z_v, $v = 1, \ldots, V$ be arbitrary, spatially indexed random variables, and suppose $\mathrm{E}\,z_v = \mathrm{E}\,z_{v'}$ for all v, v'. The population variogram is defined as

$$\nu(z_v, z_{v'}) = \mathrm{E}\,(z_v - z_{v'})^2.$$

Then note that

$$\nu(z_v, z_{v'}) = \operatorname{Var} z_v + \operatorname{Var} z_{v'} - 2\operatorname{Cov}(z_v, z_{v'}).$$

Under stationarity, $\operatorname{Var} z_v = \operatorname{Var} z_{v'}$, and hence $\nu(z_{v_i}, z_{v_j}) = \nu(z_{v_k}, z_{v_l})$ for $\|v_i - v_j\| = \|v_k - v_l\|$. In our model, we have $\mathrm{E}\,\boldsymbol{d}_{nv} = \mathrm{E}\,\boldsymbol{d}_{nv'}$, so one could consider a variogram defined by

$$\mathrm{E}\,(d_{nvq} - d_{nv'q})^2$$

which is equivalent to

$$\operatorname{Var} d_{nv} - 2\operatorname{Cov}(d_{nvq}, d_{nv'q}) + \operatorname{Var} d_{nv'}.$$

However, d_{nvq}, $v = 1, \ldots, V$ are not stationary due to the spatially-referenced temporal effects (i.e., $\xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n$). It will turn out that is easier to use the *covariogram* to estimate the spatial correlation in the presence of other variance components, so long as we use statistics in the empirical covariogram such that the other variance components are constant across spatial lags.

Specifically, we fit the following function:

$$\delta(z_v, z_{v'}) = \begin{cases} \lambda_0 + \lambda_1 & v = v', \\ \lambda_0 + \lambda_1 e^{-\theta\|v - v'\|} & v \neq v'. \end{cases} \tag{3.20}$$

That is, we have replaced the nugget effect with a "bias term" that is present for all distances. This allows us to estimate spatial correlation in the presence of subject-specific and spatially varying micro-scale variance owing to $\xi_{nv}^2 \mathbf{K}_n \boldsymbol{\Psi}_{nv} \mathbf{K}_n$. Note that in our model, $\lambda_1 = \sigma_{b_q}^2$; however, we use the estimator defined in the previous section to estimate the variance component rather than the estimate that can be obtained from the estimated covariogram. A generic form for the empirical covariogram is

$$\hat{\delta}(h) = \frac{1}{N_h} \sum_{\{v, v'\}:\ \|v - v'\| \in (h - \nu, h + \nu]} z_v z_{v'} \tag{3.21}$$

where 2ν represents the bin width. The estimator we use is described below.

Estimating the spatial correlation of the vertex random effect

It seems biologically reasonable to assume that the correlation structure in the vertex random effects is equivalent to the spatial correlation in the subject-vertex interaction. Moreover, the estimation of the exponential covariogram correlation parameter from a single realization of a random field can be highly unreliable. Finally, this assumption enables more accurate estimation of the correlation parameter.

Estimating the spatial correlation of the subject-vertex interaction effect

Under the assumption that $\mathbf{\Gamma}_{v,v'} = \mathbf{\Omega}_{v,v'}$, we have

$$\begin{aligned} \mathrm{E}\left(\boldsymbol{d}_{nv}\boldsymbol{d}'_{nv'}\right) &= \mathrm{Cov}\left(\boldsymbol{d}_{nv}, \boldsymbol{d}_{nv'}\right) + \boldsymbol{\beta}\boldsymbol{\beta}' \\ &= \mathbf{\Gamma}_{v,v'}\mathbf{U} + \mathbf{S} + \mathbf{\Omega}_{v,v'}\mathbf{B} + \boldsymbol{\beta}\boldsymbol{\beta}' \\ &= \mathbf{\Omega}_{v,v'}(\mathbf{U} + \mathbf{B}) + \boldsymbol{\beta}\boldsymbol{\beta}' + \mathbf{S} \end{aligned}$$

Then we can parameterize (3.20) in terms of our model:

$$\delta_q(d_{nvq}, d_{nv'q}) = \begin{cases} \sigma^2_{u_q} + \sigma^2_{s_q} + \sigma^2_{b_q} + \beta^2_q & v = v', \\ (\sigma^2_{s_q} + \beta^2_q) + (\sigma^2_{u_q} + \sigma^2_{b_q})e^{-\theta_q d(v,v')} & v \neq v'. \end{cases} \tag{3.22}$$

To estimate θ_q, we can use (3.21) for the product $d_{nvq}d_{nv'q}$, $v \neq v'$. We calculated an empirical covariogram for each subject, task, and parcel. We used fifteen equally sized bins to a maximum lag distance of one half the maximum distance between vertices in a given parcel. The empirical covariogram for each lag distance bin was averaged across subjects, and the average of the distance between pairs of observations falling into a given lag distance bin was calculated (this could be somewhat different from the midpoint of the lag distance bin). Then the covariogram parameters were estimated from the mean empirical covariogram by minimizing the squared errors of the mean empirical

covariogram for each lag distance bin and the theoretical covariogram evaluated at these averages.

3.3.4 BLUEs and BLUPs

Let us consider estimators for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ when the covariance matrix is known. In practice, we will substitute the estimators developed in the previous sections. We will use Henderson's equations to define the best linear unbiased predictors (BLUPs) for the random effects and the best linear unbiased estimators (BLUEs) for the fixed effects (e.g., McCulloch et al. 2008). The GLS estimator can be written as

$$\hat{\boldsymbol{\beta}} = \left\{(1'_N \otimes 1'_V \otimes \mathbf{I}_Q)\boldsymbol{\Sigma}^{-1}(1_N \otimes 1_V \otimes \mathbf{I}_Q)\right\}^{-1} (1'_N \otimes 1'_V \otimes \mathbf{I}_Q)\boldsymbol{\Sigma}^{-1}\boldsymbol{d} \tag{3.23}$$

and we have

$$\operatorname{Cov}\hat{\boldsymbol{\beta}} = \left\{(1'_N \otimes 1'_V \otimes \mathbf{I}_Q)\boldsymbol{\Sigma}^{-1}(1_N \otimes 1_V \otimes \mathbf{I}_Q)\right\}^{-1}. \tag{3.24}$$

The BLUP for $\boldsymbol{u}$ is equal to the expected value of the conditional distribution of $\boldsymbol{u}$ given the data. Using the form that appears on p.315 of McCulloch et al. 2008:

$$\hat{\boldsymbol{u}} = \boldsymbol{\Gamma}(1'_N \otimes \mathbf{I}_V \otimes \mathbf{I}_Q)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{d} - 1_N \otimes 1_V \otimes \hat{\boldsymbol{\beta}}\right). \tag{3.25}$$

Now consider the covariance of $\hat{\boldsymbol{u}}$:

$$\operatorname{Cov}\hat{\boldsymbol{u}} = \boldsymbol{\Gamma}(1'_N \otimes \mathbf{I}_V \otimes \mathbf{I}_Q)\boldsymbol{\Sigma}^{-1}\operatorname{Cov}\left\{\boldsymbol{d} - (1_N \otimes 1_V \otimes \mathbf{I}_Q)\hat{\boldsymbol{\beta}}\right\}\boldsymbol{\Sigma}^{-1}(1_N \otimes \mathbf{I}_V \otimes \mathbf{I}_Q)\boldsymbol{\Gamma},$$

which involves the quantity,

$$\begin{aligned}\operatorname{Cov}(\boldsymbol{d}, \hat{\boldsymbol{\beta}}) &= (\operatorname{Cov}\boldsymbol{d})\,\boldsymbol{\Sigma}^{-1}\,(1_N \otimes 1_V \otimes \mathbf{I}_Q)\left\{(1'_N \otimes 1'_V \otimes \mathbf{I}_Q)\boldsymbol{\Sigma}^{-1}(1_N \otimes 1_V \otimes \mathbf{I}_Q)\right\}^{-1} \\ &= (1_N \otimes 1_V \otimes \mathbf{I}_Q)\operatorname{Cov}\hat{\boldsymbol{\beta}}. \end{aligned} \tag{3.26}$$

Then we have

$$\begin{aligned}
&\operatorname{Cov} \hat{\boldsymbol{u}} = \\
&= \boldsymbol{\Gamma}(1_N' \otimes \mathbf{I}_V \otimes \mathbf{I}_Q)\boldsymbol{\Sigma}^{-1} \{\boldsymbol{\Sigma} - \\
&\quad (1_N \otimes 1_V \otimes \mathbf{I}_Q) \left(\operatorname{Cov} \hat{\boldsymbol{\beta}}\right) (1_N' \otimes 1_V' \otimes \mathbf{I}_Q)\Big\} \boldsymbol{\Sigma}^{-1}(1_N \otimes \mathbf{I}_V \otimes \mathbf{I}_Q)\boldsymbol{\Gamma}. \\
&= \boldsymbol{\Gamma} \left(1_N' \otimes \mathbf{I}_V \otimes \mathbf{I}_Q\right) \boldsymbol{\Sigma}^{-1} \{\mathbf{I}_{NVQ} - \\
&(1_N \otimes 1_V \otimes \mathbf{I}_Q) \left(\operatorname{Cov} \hat{\boldsymbol{\beta}}\right) (1_N' \otimes 1_V' \otimes \mathbf{I}_Q) \boldsymbol{\Sigma}^{-1}\Big\} (1_N \otimes \mathbf{I}_V \otimes \mathbf{I}_Q)\boldsymbol{\Gamma}. \qquad (3.27)
\end{aligned}$$

We will first describe a method to invert $\boldsymbol{\Sigma}$, which is necessary to calculate the BLUEs and BLUPs. Due to the size of the covariance matrices, direct inversion of $\boldsymbol{\Sigma}$ is difficult at best. We now describe some computational tricks.

Consider the components of the covariance matrix wherein the subjects are independent:

$$\mathbf{F}_n = \mathbf{J}_V \otimes \mathbf{S} + \boldsymbol{\Omega} + \oplus_{v=1}^{V} \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n.$$

and

$$\mathbf{F} = \mathbf{I}_N \otimes \mathbf{J}_V \otimes \mathbf{S} + \mathbf{I}_N \otimes \boldsymbol{\Omega} + \oplus_{n=1}^{N} \oplus_{v=1}^{V} \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n.$$

We can divide $\mathbf{F}$ it into N submatrices which each require the inversion of a $VQ \times VQ$ matrix.

Now our goal is to invert

$$\mathbf{J}_N \otimes \boldsymbol{\Gamma} + \mathbf{F}.$$

Since the first matrix has rank VQ, we can utilize the Sherman-Morrison-Woodbury formula to break the inversion of the $NVQ \times NVQ$ covariance matrix into the inversion of $VQ \times VQ$ matrices. Note that

$$\mathbf{J}_N \otimes \boldsymbol{\Gamma} = (1_N \otimes \mathbf{I}_{VQ})\boldsymbol{\Gamma}(1_N' \otimes \mathbf{I}_{VQ}),$$

so we have

$$\{\mathbf{F} + (\mathbf{1}_N \otimes \mathbf{I}_{VQ})\boldsymbol{\Gamma}(\mathbf{1}_N' \otimes \mathbf{I}_{VQ})\} =$$
$$\mathbf{F}^{-1} - \mathbf{F}^{-1}(\mathbf{1}_N \otimes \mathbf{I}_{VQ})\left\{\boldsymbol{\Gamma}^{-1} + (\mathbf{1}_N' \otimes \mathbf{I}_{VQ})\mathbf{F}^{-1}(\mathbf{1}_N \otimes \mathbf{I}_{VQ})\right\}^{-1}(\mathbf{1}_N' \otimes \mathbf{I}_{VQ})\mathbf{F}^{-1}.$$

Note that

$$\mathbf{F}^{-1}(\mathbf{1}_N \otimes \mathbf{I}_{VQ}) = \{_c\ \mathbf{F}_n^{-1}\}_{n=1}^N$$

and

$$(\mathbf{1}_N' \otimes \mathbf{I}_{VQ})\mathbf{F}^{-1}(\mathbf{1}_N \otimes \mathbf{I}_{VQ}) = \sum_{n=1}^N \mathbf{F}_n^{-1}$$

so the following form can be used for computationally lower cost implementation:

$$\boldsymbol{\Sigma}^{-1} = \oplus_{n=1}^N \mathbf{F}_n - \{_c\ \mathbf{F}_n^{-1}\}_{n=1}^N \left(\boldsymbol{\Gamma}^{-1} + \sum_{n=1}^N \mathbf{F}_n^{-1}\right)^{-1} \{_r\ \mathbf{F}_n^{-1}\}_{n=1}^N.$$

In practice, we replace the true covariance components with their estimators. It should be noted that the resulting estimators for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$ are in fact no longer the BLUE or BLUP, respectively. In the mixed effects literature, BLUPs derived using the estimated variance components are sometimes called eBLUPs.

3.3.5 Generalized t-test for inference in the STMM model

We are interested in calculating an approximate t-statistic for each vertex and for any arbitrary set of vertices forming an ROI. For clarity, we now introduce the index for region in this section, letting $\hat{\boldsymbol{\beta}}^{(r)}$ denote the fixed effects estimates for the rth region, $r = 1, \ldots, R$, and similarly define $\hat{\boldsymbol{u}}^{(r)}$ and $\widehat{\boldsymbol{\Sigma}}^{(r)}$. Now let V equal the *total* number of vertices across all regions. We now let $\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}^{(1)'}, \ldots, \hat{\boldsymbol{\beta}}^{(R)'}]'$, such that $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{RQ}$.

Now consider a framework for conducting inference on linear combinations of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$. BLUPs can be derived from an empirical Bayes perspective in which the prior

mean is equal to zero and the prior covariance is estimated from the data. In our context, we write the posterior distribution as

$$[\boldsymbol{u}|\boldsymbol{d}, \widehat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\beta}}] \sim \mathcal{N}\Big(\widehat{\boldsymbol{\Gamma}}\{1'_N \otimes \mathbf{I}_V \otimes \mathbf{I}_Q\}\widehat{\boldsymbol{\Sigma}}^{-1}\Big(\boldsymbol{d} - 1_N \otimes 1_V \otimes \hat{\boldsymbol{\beta}}\Big),$$
$$\widehat{\boldsymbol{\Gamma}} - \widehat{\boldsymbol{\Gamma}}\{1'_N \otimes \mathbf{I}_V \otimes \mathbf{I}_Q\}\widehat{\boldsymbol{\Sigma}}^{-1}\{1_N \otimes \mathbf{I}_V \otimes \mathbf{I}_Q\}\widehat{\boldsymbol{\Gamma}}\Big).$$

In this context, $\hat{\boldsymbol{u}} = \mathrm{E}(\boldsymbol{u}|\boldsymbol{d}, \widehat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\beta}})$. Here, the (transformed) data, $\boldsymbol{d}$, includes the realizations of the locations at which the random vertex effects were observed.

Now for an arbitrary contrast vector $\mathbf{c}$, we propose testing hypotheses of the form

$$H_0 : \mathbf{c}'[\boldsymbol{\beta}', \mathrm{E}(\boldsymbol{u}|\boldsymbol{d})']' = 0$$
$$H_A : \mathbf{c}'[\boldsymbol{\beta}', \mathrm{E}(\boldsymbol{u}|\boldsymbol{d})']' \neq 0.$$

We use $\hat{\boldsymbol{u}}$ as our estimate of $\mathrm{E}(\boldsymbol{u}|\boldsymbol{d})$. Then consider the following test statistic:

$$t = \frac{\mathbf{c}'[\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{u}}']'}{\left\{\mathbf{c}'(\widehat{\mathrm{Cov}}\,\hat{\boldsymbol{\beta}} \oplus \widehat{\mathrm{Cov}}\,\hat{\boldsymbol{u}})\mathbf{c}\right\}^{1/2}}. \tag{3.28}$$

To simplify the calculation of the test statistic, we show that $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{\beta}}$ are independent (for known covariance). First,

$$\mathrm{Cov}(\hat{\boldsymbol{u}}, \hat{\boldsymbol{\beta}}) = \boldsymbol{\Gamma}(1'_N \otimes \mathbf{I}_V \otimes \mathbf{I}_Q)\boldsymbol{\Sigma}^{-1}\mathrm{Cov}\left[\left\{\boldsymbol{d} - (\mathbf{1}_N \otimes \mathbf{1}_V \otimes \mathbf{I}_Q)\hat{\boldsymbol{\beta}}\right\}, \hat{\boldsymbol{\beta}}\right].$$

From the result in (3.26), it follows that

$$\mathrm{Cov}(\hat{\boldsymbol{u}}, \hat{\boldsymbol{\beta}}) = 0.$$

We would like to approximate the distribution of this test statistic under the null hypothesis. In the results that appear in Section 3.4.2, we conservatively use a t-distribution with $N - 1$ degrees of freedom. A method to approximate the degrees of freedom is proposed in C.5, although we do not evaluate it here.

3.4 Simulation studies

3.4.1 Assessing the accuracy of the STMM estimators

In this section, we diagnose the accuracy of the estimators presented above.

We simulated surface fMRI data using thirty subject-specific design matrices from the ToM study. The construction of the design matrices is described in Section 3.5.1. We chose the first thirty subjects in the HCP data sampler (subject IDs 100307 to 124422). From these design matrices, we chose two covariates of interest (xMental and xRandom) and ten nuisance covariates (intercept, indicator variable for session, and the eight covariates forming the piecewise linear spline basis). We used the locations of vertices corresponding to parcel 82 in our version of the Gordon networks. This parcel was chosen because it contained an average number of vertices (215, where the average of all parcels is 221 vertices).

There were $30 \times 215 \times (10 + 3 + 1) = 90,300$ parameters estimated in the first-level of the model (10 nuisance covariates, 3 AR parameters, and one error variance for each subject and each vertex), which were estimated from a total of 30×215×548 =3,534,600 observations. To ensure realistic values, the coefficients for the ten nuisance covariates were set equal to estimates from the first-level analyses of these subjects' data. The innovation variance, which we denote as τ^2_{nv}, was equal to 30,000 for all vertices and all subjects, which is similar to estimates from the HCP data.

In Tables 3.1 and 3.2, the AR parameters were equal to 0.3, -0.15, and 0.1 for all vertices and all subjects. In Table 3.3, the AR parameters were equal to 0.2, 0.1, and 0. In Tables 3.1 and 3.3, the spatial dependence parameters were 0.2 and 0.2, while in Table 3.2, they were 0.5 and 0.1. These values are within the range of those calculated

	Truth	df	Mean STMM	Var STMM	Bias^2 STMM	MSE STMM
τ^2_{nv}	30000	-	29815	3443178	34171	3477349
ϕ_{nv1}	0.3000	-	0.2984	0.0021	0.0000	0.0021
ϕ_{nv2}	-0.1500	-	-0.1514	0.0020	0.0000	0.0020
ϕ_{nv3}	0.1000	-	0.0985	0.0020	0.0000	0.0020
β_1	20.0	-	20.3	18.1	0.1	18.2
$\sigma^2_{u_1}$	75.0	214	75.5	624.9	0.3	625.2
$\sigma^2_{s_1}$	250	29	231	8334	368	8703
$\sigma^2_{b_1}$	750	6206	754	3836	20	3856
θ_{b_1}	0.200	-	0.205	0.002	0.000	0.002
β_2	-20.0	-	-20.0	13.4	0.0	13.4
$\sigma^2_{u_2}$	75.0	214	75.8	844.1	0.7	844.8
$\sigma^2_{s_2}$	250	29	241	7894	79	7974
$\sigma^2_{b_2}$	750	6206	754	4147	14	4161
θ_{b_2}	0.200	-	0.205	0.002	0.000	0.003

Table 3.1: Accuracy of estimators for 100 simulations with 30 subjects each in which 215 vertices were located according to a Gordon Parcel (ID 82). Note that τ^2_{nv} corresponds to the innovation variance for the AR(3) process.

from the HCP data.

The average behavior of the estimators is very close to their true values. In all scenarios, τ^2_{nv}, ϕ_{nv1}, ϕ_{nv2}, and ϕ_{nv3} were slightly underestimated. This is due to approximating the covariance matrix with the first twenty lags. However, the amount of bias in the bias-reduced parameters is negligible and has little effect on the other parameter estimates. Note that the asymptotic variance for a sample partial autocorrelation parameter for white noise data is $1/T$. The variance of the biased-reduced estimators of the AR coefficients is quite similar, where here, $T = 548$.

	Truth	df	Mean STMM	Var STMM	Bias^2 STMM	MSE STMM
τ^2_{nv}	30000	-	29815	3421582	34041	3455622
ϕ_{nv1}	0.3000	-	0.2984	0.0020	0.0000	0.0021
ϕ_{nv2}	-0.1500	-	-0.1515	0.0020	0.0000	0.0020
ϕ_{nv3}	0.1000	-	0.0984	0.0020	0.0000	0.0020
β_1	20.0	-	20.7	12.4	0.5	13.0
$\sigma^2_{u_1}$	75.0	214	77.1	164.4	4.5	169.0
$\sigma^2_{s_1}$	250	29	238	5455	148	5603
$\sigma^2_{b_1}$	750	6206	749	888	1	889
θ_{b_1}	0.500	-	0.529	0.020	0.001	0.021
β_2	-20.0	-	-20.2	14.6	0.0	14.7
$\sigma^2_{u_2}$	75.0	214	78.0	1980.1	9.2	1989.2
$\sigma^2_{s_2}$	250	29	233	20773	300	21073
$\sigma^2_{b_2}$	750	6206	770	26942	407	27349
θ_{b_2}	0.100	-	0.111	0.002	0.000	0.002

Table 3.2: Accuracy of estimators for 100 simulations with 30 subjects each in which 215 vertices were located according to a Gordon Parcel (ID 82). In this scenario, the dependence parameters θ_{b_1} and θ_{b_2} differ from Table 3.1.

	Truth	df	Mean STMM	Var STMM	Bias^2 STMM	MSE STMM
τ^2_{nv}	30000	-	29811	3434698	35882	3470580
ϕ_{nv1}	0.2000	-	0.1988	0.0020	0.0000	0.0020
ϕ_{nv2}	0.1000	-	0.0969	0.0020	0.0000	0.0021
ϕ_{nv3}	0.0000	-	-0.0008	0.0020	0.0000	0.0020
β_1	20.0	-	19.9	16.2	0.0	16.2
$\sigma^2_{u_1}$	75.0	214	77.0	776.0	4.1	780.2
$\sigma^2_{s_1}$	250	29	241	8521	84	8605
$\sigma^2_{b_1}$	750	6206	750	3517	0	3517
θ_{b_1}	0.200	-	0.201	0.002	0.000	0.002
β_2	-20.0	-	-20.0	17.5	0.0	17.5
$\sigma^2_{u_2}$	75.0	214	75.9	808.2	0.9	809.0
$\sigma^2_{s_2}$	250	29	244	7947	38	7985
$\sigma^2_{b_2}$	750	6206	749	3747	2	3749
θ_{b_2}	0.200	-	0.203	0.002	0.000	0.002

Table 3.3: Accuracy of estimators for 100 simulations with 30 subjects each in which 215 vertices were located according to a Gordon Parcel (ID 82). In this scenario, the AR coefficients differ from Table 3.1.

3.4.2 Comparing type-1 error rates and power in the MUMM, ROIMM, and STMM

We compared the performance of the MUMM, ROIMM, and STMM using simulations from the subject design matrices from the first thirty subjects of the ToM and the distance structure from parcel 82 (215 vertices). The following parameters were constant across all simulations and scenarios: $\beta_1 = 20$; $\beta_2 = 0$; $\sigma^2_{s_1} = \sigma^2_{s_2} = 250$; $\sigma^2_{b_1} = \sigma^2_{b_2} = 750$; τ^2_{nv}, ϕ_{nv1}, ϕ_{nv2}, and ϕ_{nv3} were set equal to estimates from the first-level analysis using the biased-reduced estimators.

We evaluated four scenarios to examine the effect of spatial dependence and vertex random effects on power (sensitivity) and type-1 error rates (specificity). In the first scenario, we set the dependence parameters for both the xMental and xRandom to 5 (for both the subject-vertex and vertex random effects). This is approximately equivalent to zero spatial dependence (the smallest distance is 1.9 mm, corresponding to a correlation less than 0.0001). We also set the vertex variance component for both xMental and xRandom to zero. In the second scenario, we kept the dependence parameter equal to 5, but set the vertex variance components equal to 100. In the third scenario, we set the dependence parameters equal to 0.2 but the vertex variance components equal to 0. This corresponds to a spatial mixed model with subject-vertex spatial random effects but no vertex random effects. In the final scenario, we let the dependence parameters equal 0.2 and the vertex variance components equal 100. We conducted 300 simulations for each scenario and set the $\alpha-$level equal to 0.05 with critical value from a t-distribution with $N - 1$ (here, 29) degrees of freedom. For the region-level t-statistic for the MUMM, the degrees of freedom under the independence assumption were equal to V (here, 215) since the variance estimate is an average of V (assumed to be independent) chi-squared variables.

Examining the region-level inference in the three models, all methods have nearly perfect power for xMental under the parameter values in the simulations (Table 3.4). With respect to the type-1 error rates, MUMM has drastically inflated type-1 error rates for region-level inference for all scenarios. The error rates were inflated even when there was no spatial dependence in the subject-vertex random effects and no vertex random effects because the subject-specific random effects, $\boldsymbol{s}_n$, induce a constant correlation between all vertices. The model simulated in scenario 1 is similar to the models in Derado et al. (2010) and Bowman (2005). In scenario 1, the type 1 error rate in the ROIMM and STMM is approximately equal to the nominal rate, and both methods had nearly perfect power for the contrast. In Scenarios 2 and 3, both ROIMM and STMM preserve the nominal α-level with nearly perfect power. In Scenario 4, however, the type-1 error rate of ROIMM is inflated, nearly three times the nominal rate, while the STMM is slightly conservative. The power for ROIMM is higher than STMM, but at the cost of an unacceptably high type-1 error rate. Thus ROIMM accounts for spatial correlation due to the subject and subject-vertex random effects, but is inappropriate to use when the data follow a model with vertex random effects.

Examining the vertex-level inference, STMM is much more powerful than MUMM in Scenarios 1 and 3 when the vertex random effects are equal to zero, but we see a large decrease in power in the STMM when there are vertex random effects. For $\beta_1 + u_{v1}$, the power of the STMM is greater than MUMM in Scenarios 1 and 3, but is more similar when there exist vertex random effects without spatial dependence, and the powers are approximately equal when there exist vertex random effects with spatial dependence. With respect to $\beta_2 + u_{v2}$, Scenarios 1 and 3 examine type one error rates since $\sigma^2_{u_1} = \sigma^2_{u_2} = 0$ implies $\mathrm{E}\,(u_{v2}|\boldsymbol{d})$ is equal to zero. We see that STMM is overly conservative, while the type one error rate of MUMM is approximately equal to the nominal α-level. In Scenarios 2 and 4, the rates for $\beta_2 + u_{v2}$ represent the power to detect the conditional

		Region			Vertex		
Scenario	Model	$\beta_1 = 20$ (Power)	$\beta_2 = 0$ (Type 1)	Contrast (Power)	$\beta_1 + u_{v1}$ (Power)	$\beta_2 + u_{v2}$ (Type 1*)	Contrast (Power)
$\theta_1 = \theta_2 = 5$ $\sigma^2_{u_1} = \sigma^2_{u_2} = 0$	MUMM	1.00	0.72	1.00	0.61	0.05	0.38
	ROIMM	1.00	0.05	0.99	-	-	-
	STMM	1.00	0.06	0.99	1.00	0.02	0.99
$\theta_1 = \theta_2 = 5$ $\sigma^2_{u_1} = \sigma^2_{u_2} = 100$	MUMM	1.00	0.70	1.00	0.57	0.19	0.41
	ROIMM	1.00	0.05	1.00	-	-	-
	STMM	1.00	0.06	1.00	0.63	0.02	0.35
$\theta_1 = \theta_2 = 0.2$ $\sigma^2_{u_1} = \sigma^2_{u_2} = 0$	MUMM	1.00	0.72	1.00	0.61	0.05	0.39
	ROIMM	1.00	0.05	0.99	-	-	-
	STMM	1.00	0.04	0.99	1.00	0.02	0.98
$\theta_1 = \theta_2 = 0.2$ $\sigma^2_{u_1} = \sigma^2_{u_2} = 100$	MUMM	1.00	0.79	1.00	0.56	0.19	0.42
	ROIMM	1.00	0.14	0.95	-	-	-
	STMM	0.99	0.03	0.86	0.57	0.02	0.31

Table 3.4: Power and type 1 error rates for estimated main effects, their contrast, main effects plus predicted random effects, and their contrast based on 300 simulations for each scenario. Note that the scenarios represent (1) approximately zero spatial correlation in the subject-vertex random effects and no vertex random effects; (2) approximately zero correlation with vertex random effects; (3) spatial correlation in the subject-vertex random effects with no vertex random effects; and (4) spatial correlation with vertex random effects. *Note that in vertex-specific inference on $\beta_2 + u_{v2}$, the proportions represent type 1 error rates when $\sigma^2_{u_1} = \sigma^2_{u_2} = 0$; however, when $\sigma^2_{u_1} = \sigma^2_{u_2} = 100$, these represent the power to detect the conditional mean when the unconditional mean is equal to zero.

mean, $\mathrm{E}(u_{v2}|\boldsymbol{d})$, when the marginal mean is equal to zero. MUMM detects this small signal approximately 20% of the time, while STMM does not detect a signal. For the contrast, STMM is much more powerful than MUMM in Scenarios 1 and 3 with nearly perfect power but less powerful in Scenarios 2 and 4. The power of MUMM to detect the contrast is similar across all scenarios, which contrasts with STMM.

3.5 Analysis of ToM HCP data

3.5.1 Motivating dataset

We applied the MUMM, ROIMM, and STMM to data from a social cognition / theory of mind experiment of the MGH-UCLA Human Connectome Project (HCP). Theory of mind refers to the ability to intuit another person's actions or feelings. In the HCP experiment, subjects in an fMRI scanner viewed cartoons that either depicted shapes acting in human-like ways (e.g., a large triangle leading a smaller triangle out of a maze) or in random ways, which were the "mentalizing" (hereafter, xMental) and "random" (xRandom) tasks, respectively. For details of the experimental paradigm see Barch et al. (2013). Whole-brain data were acquired from two sessions with 274 volumes each using gradient-echo EPI with an eight-band multifactor approach and $2 \times 2 \times 2$ mm voxels (repetition time (TR) = 720 ms; echo time (TE) = 33.1 ms; flip angle=52°; field of view = 208×180 mm (readout $\times$ phase-encoding); acquisition matrix = 104×90; slice thickness = 2.0 mm). We used the minimally pre-processed data from the first twenty subjects of the unrelated 100 data sampler released August 5, 2014. The minimally preprocessed data include fMRI data registered to the Freesurfer 32k spherical template, the end result of which is a set of approximately 30,000 time series for each cortex and each session and each subject on a standard mesh where the vertex numbers correspond to spatially matched locations (Glasser et al., 2013). The data do contain a small amount of smoothing (2 mm on the surface), but we elected to use these minimally preprocessed data rather than process our own completely unsmoothed data.

The HCP project (Principal Investigators: Bruce Rosen, M.D., Ph.D., Martinos Center at Massachusetts General Hospital; Arthur W. Toga, Ph.D., University of California, Los Angeles, Van J. Weeden, MD, Martinos Center at Massachusetts General Hospital)

is supported by the National Institute of Dental and Craniofacial Research (NIDCR), the National Institute of Mental Health (NIMH), and the National Institute of Neurological Disorders and Stroke (NINDS). Collectively, the HCP is the result of efforts of co-investigators from the University of California, Los Angeles, Martinos Center for Biomedical Imaging at Massachusetts General Hospital (MGH), Washington University in St. Louis, and the University of Minnesota.

We modeled the right cerebral cortex, which contains 29,716 vertices (the 32k Freesurfer template excluding the medial wall). Each session contains 274 time points with one session phase-coded RL and the other LR. We concatenated the two sessions and assume the AR errors between the two sessions are independent but generated from processes with the same AR parameters. We included an indicator variable for session to account for differences in the mean image owing to different phase codings. When estimating the AR parameters, we modified the calculation of the sample autocorrelation to exclude pairs of observations from different sessions.

The main covariates of interest, xMental and xRandom, were generated by convolving task onsets and durations with the canonical HRF in SPM12 (Functional Imaging Laboratory). We also used SPM12 to generate four additional covariates: the derivative of xMental and xRandom with respect to the temporal delay parameter and with respect to the dispersion parameter for each task.

To estimate drift, we include separate piecewise linear splines with three evenly spaced knots (i.e., four covariates) for each session. We also included the affine registration parameters to correct for motion-induced activation. We found that the covariate effects for the motion parameters for the first session differed greatly from those for the second session, so we included the interaction between session and the motion parameters. This resulted in a total of twenty-eight covariates (see Table C.2 in the Appendix).

We defined spatially distinct regions according to the cortical parcellation in Gordon et al. (2014), which is based on correlations between the BOLD signal at each vertex in resting-state fMRI data. The Gordon parcellation for the right cerebral cortex comprises 172 spatially contiguous networks defined in the Freesurfer 32k template. These networks range in size from ten to 829 vertices. The boundaries between networks are not classified. A total of 8,509 out of 29,716 vertices are unclassified. Each unclassified vertex was assigned to the parcel containing the closest classified vertex. When a vertex was equidistant from classified vertices in two different parcels, then the vertex was assigned based on which parcel contained the second closest vertex. This resulted in a unique classification of all vertices. The revised parcels range in size from 29 to 986 vertices.

We calculated the geodesic distance between vertices in the Freesurfer 32k spherical template (fsaverage6) using Connectome Workbench. One could consider subject-specific distances based on subjects' mid-thickness cortex. However, computations simplify if we use one distance matrix for all subjects. We believe the spherical template is the most appropriate space in which to assume an equivalent spatial correlation pattern across subjects.

In general, the exponential covariogram fit the mean of the subject-specific covariograms well (Figure 3.1). For very small parcels, there were fewer data points available to calculate the empirical covariogram, which led to a non-monotonic empirical covariogram (Figure 3.2). For some of the smallest parcels, the variance components of the spatial random effects was equal to zero, indicating a homogeneous parcel with no random effects. If $\sigma^2_{u_q} = 0$, then the covariogram was not used in smoothing.

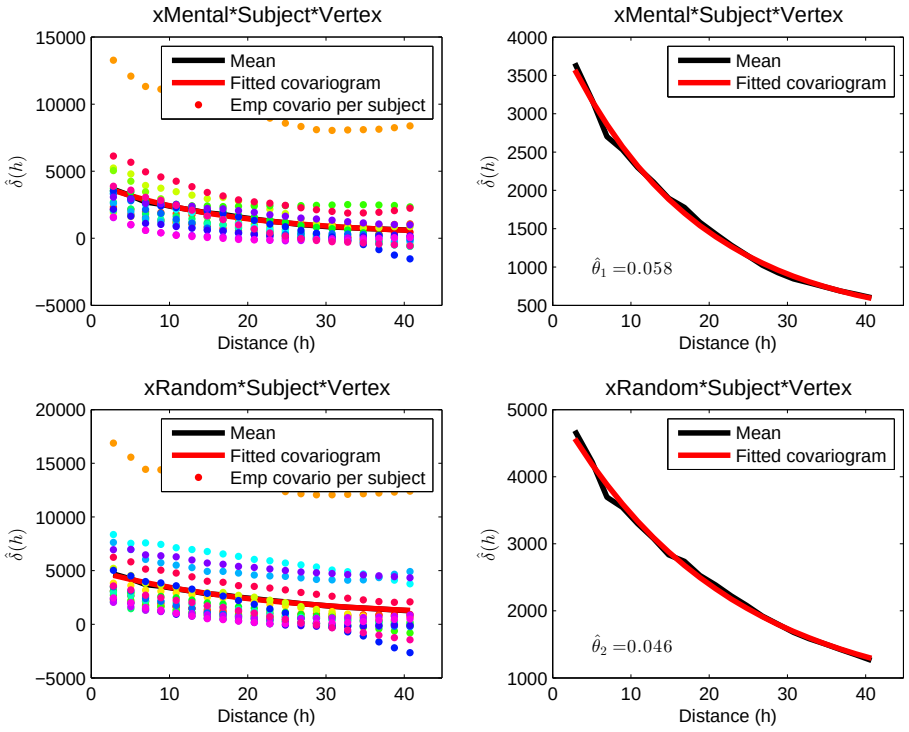


Figure 3.1: Empirical covariogram and fitted exponential covariogram as defined in Section 3.3.3 for the subject-vertex random effects of xMental and xRandom for Gordon Parcel 15, which contains 777 vertices.

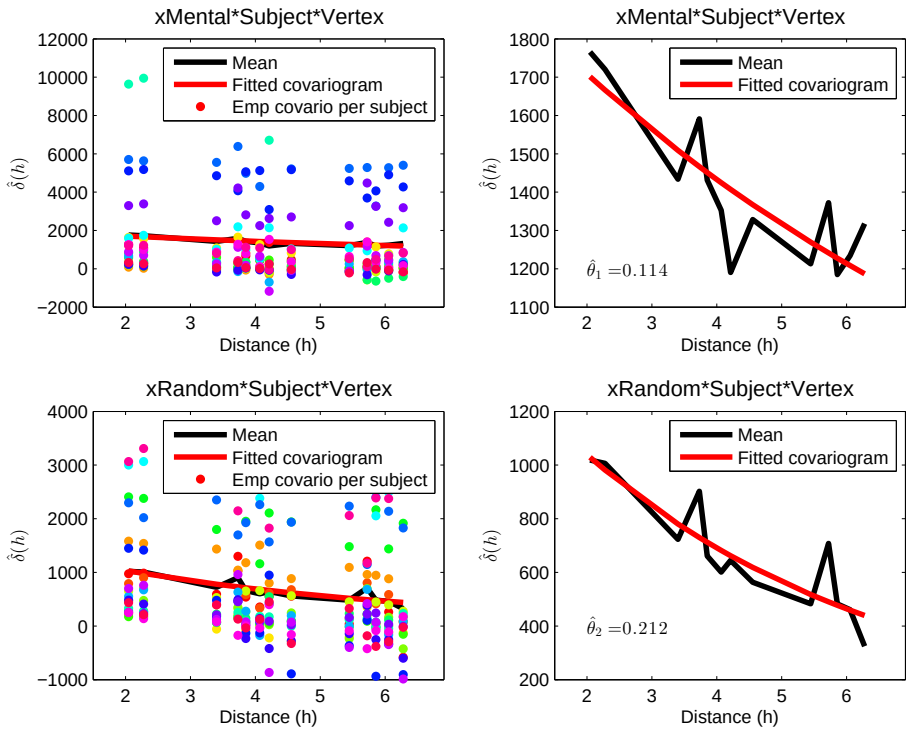


Figure 3.2: Empirical covariogram and fitted exponential covariogram for the subject-vertex random effects of xMental and xRandom for the smallest Gordon parcel (ID 169), which contains 29 vertices. This represents the worst-case scenario.

3.5.2 Results

The entire analysis of the right cerebral cortex from twenty subjects took less than one hour in Matlab on a desktop with a quad-core hyper-threaded 3.60 GHz CPU with thirty gigabytes of memory. The first-level estimation with bias-reduced AR parameters took approximately twenty minutes (one minute per subject). The second-level estimation for all twenty subjects took less than thirty minutes. This does not include the one-time cost of calculating the distances between vertices or the $\mathbf{M}_n$ matrices used in bias reduction in (3.14). In our current implementation, the number of subjects that can be analyzed is limited by the amount of memory rather than computation time. The largest parcel is 986 vertices, resulting in a $39,440 \times 39,440$ precision matrix ($986 \times 20 \times 2$), which is not sparse since we have no independence anywhere.

The smoothing of the STMM is apparent in Figure 3.3, which overall has a less speckled appearance than the MUMM. The degree of smoothing is moderate to low. Recall that the minimally preprocessed HCP data were smoothed with a FWHM 2 mm Gaussian filter, so the effect of smoothing from the STMM would likely be greater for unsmoothed data. The magnitude of the coefficients is similar in all models. Note the temporal parietal junction, superior temporal gyrus, and dorsolateral prefrontal cortex have the largest coefficients, which has also been found in other ToM studies.

The models differ markedly with respect to the t-statistic images (Figure 3.4). Overall, the STMM t-statistics tend to be lower than both the MUMM and ROIMM. This is not surprising since we found both the ROIMM and MUMM had inflated t-statistics in the simulations. However, some of the differences correspond to areas that have been associated with ToM in other experiments; for example, vertices in the temporal parietal junction have high t-statistics in the MUMM and ROIMM but the t-statistics in the STMM would not survive corrections for multiple testing. This suggests that there may

be a cost in terms of power to treating the vertices as random rather than fixed effects.

3.6 Discussion

We present a spatiotemporal mixed model for localizing brain activation from fMRI data. Our contributions are the following. First, we introduce spatial random effects that capture population activation, which leads to automated smoothing. This obviates the need for smoothing to increase the power to detect activated locations, since the amount of power is now determined by the data. Second, we utilize subject-vertex random effects to allow subject-specific deviations in activation and/or alignment. This obviates the need for smoothing to increase the overlap of features between subjects. Third, we develop a unified model that includes subject- and vertex-specific autoregressive errors, which contrasts with previous methods that use the output from a first-level analysis. Fourth, we leverage improvements in cortical registration and improvements in parcellation by using the geodesic distances between vertices within a Gordon parcel. Fifth, we develop fast estimators of spatial dependence that can be used for whole-brain studies, which improve upon previous multi-subject spatial models that assume a constant correlation between all locations within a region.

We do not address the multiple testing problem in this study. An idea for future research is to control the family-wise error rate (FWE) for a pre-specified region using a novel approach based on approximating the distribution of the maximum statistic given the estimated correlation structure of the STMM. Conditioning on the correlation estimate is somewhat unappealing but is similar to the approach used in RFT, in which the FWHM is estimated. One issue with applying the current RFT approach to our model is that we have discontinuities in our images due to the parcellation, which could render

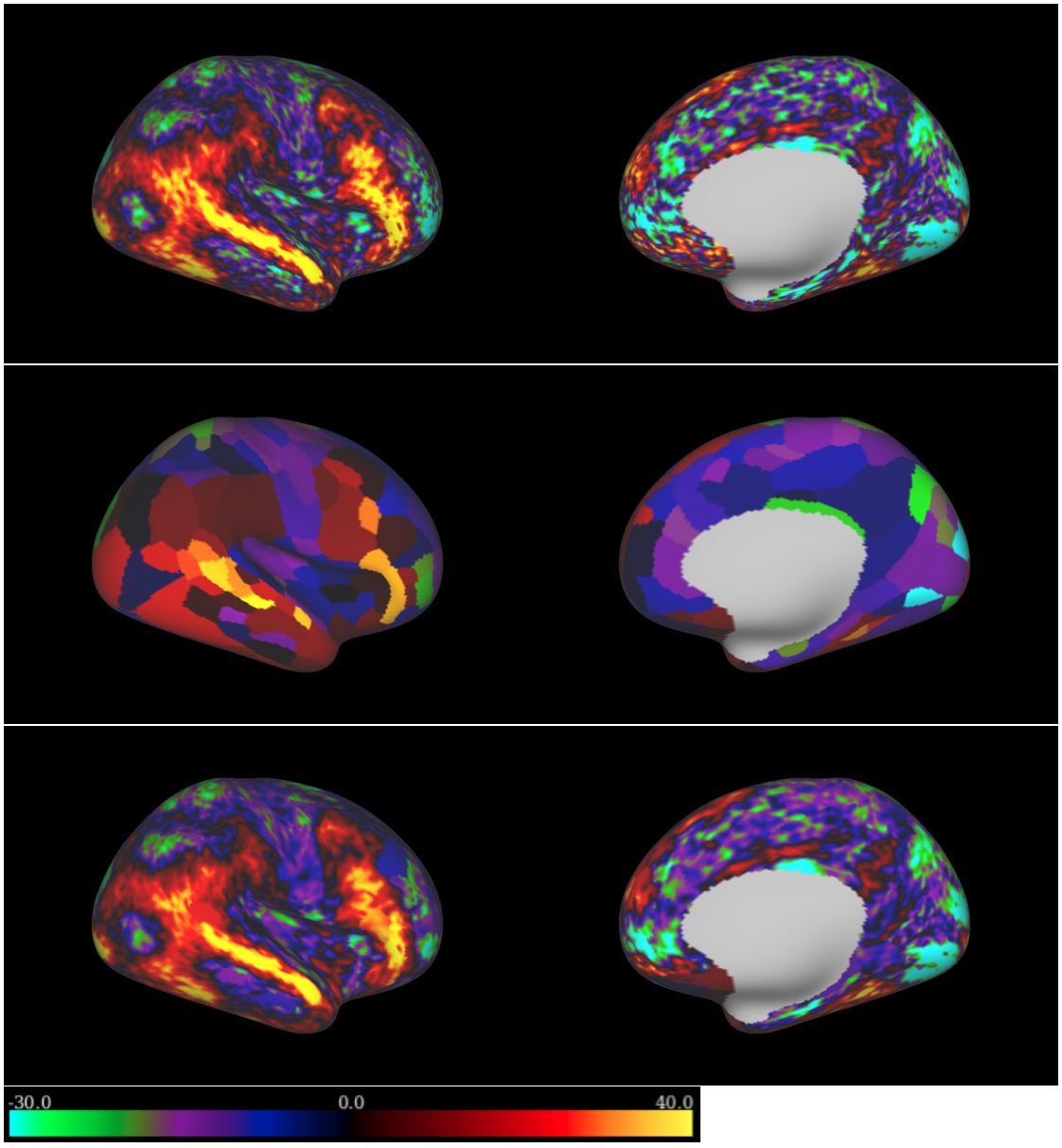


Figure 3.3: The contrast coefficient (xMental - xRandom) from the MUMM (top), ROIMM, and STMM (bottom) in the right cerebral cortex (structural dataset used to project vertices: Q1-Q6 Related 440 subjects very inflated). Values in the MUMM and ROIMM correspond to the population coefficients, while STMM coefficients represent the region-level effect plus the predicted vertex-random effect.

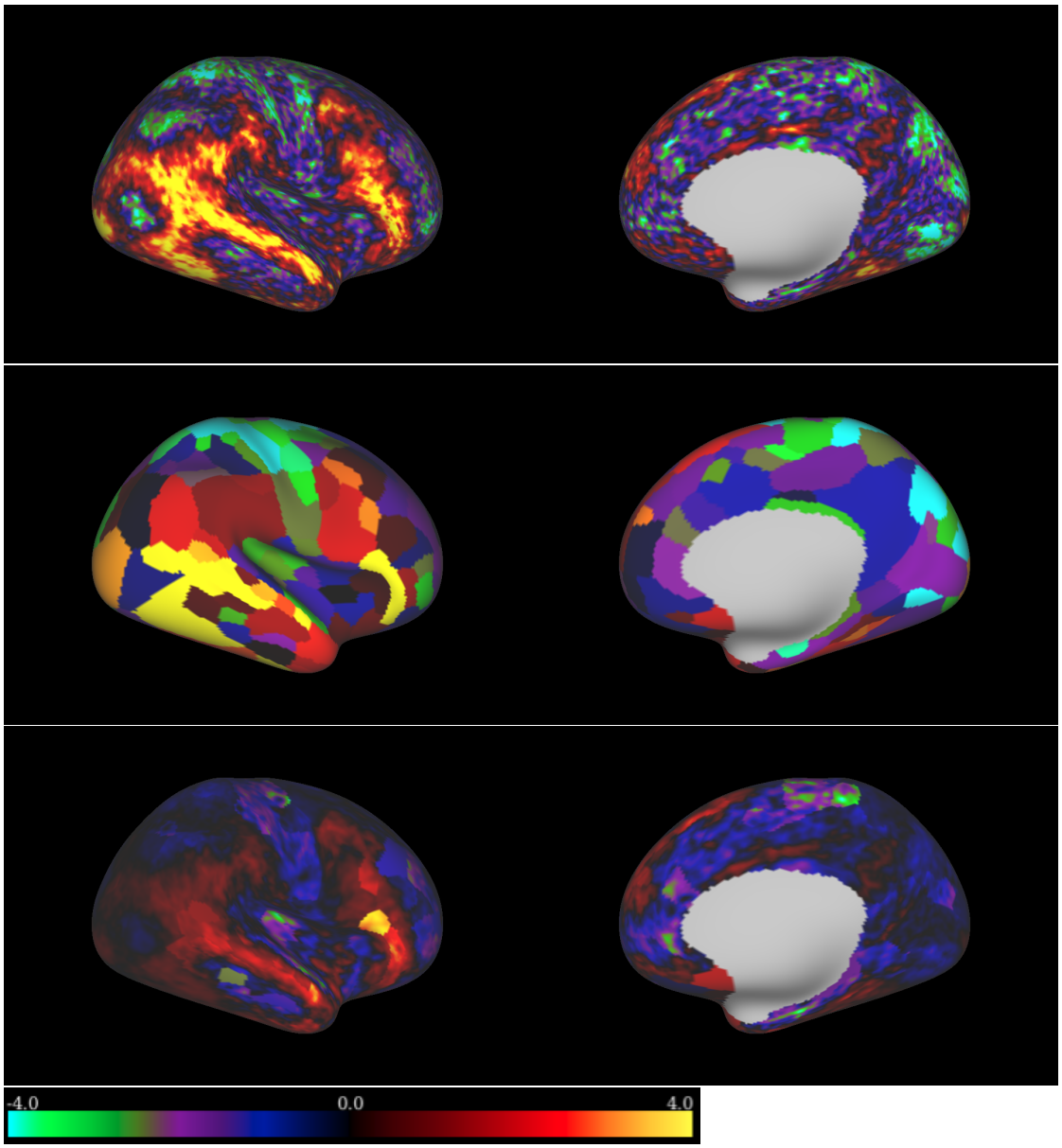


Figure 3.4: The contrast t-statistic (xMental - xRandom) from the MUMM (top), ROIMM, and STMM (bottom) in the right cerebral cortex (structural dataset used to project vertices: Q1-Q6 Related 440 subjects very inflated).

the approximations used in RFT inaccurate. If these inaccuracies are ignored, one could estimate the FWHM of a Gaussian kernel fit to images such as Figure 3.4 and determine whether the image is sufficiently smooth for the approximations used in RFT to be accurate. Then clusters could be created using the usual rules, e.g., p-values< 0.01 to threshold activated regions and cluster extent greater than 30, and critical values for the test statistics could be calculated from the estimated FWHM. Permutation-based approaches offer an alternative approach to RFT that make fewer assumptions (Nichols and Holmes, 2002). In our data application, one would multiply d_{nvq} by a randomly chosen value in $\{-1, 1\}$ for all n, v, and q and then refit the second level of the STMM. Repeating this hundreds of times generates an empirical distribution of the test statistic under the null hypothesis that there is no contrast between xMental and xRandom. Although computationally expensive, this process is somewhat feasible because our second level estimation is relatively fast, although it would require a cluster (e.g., in our application, the second level required approximately 30 minutes, so we would need 500 hours of computing time to generate 1,000 samples). The permutation test can be extended to control the family-wise error rate by generating an empirical distribution of the maximum statistic (Nichols and Hayasaka, 2003). The permutation approach has the additional advantage that it sidesteps the issue of approximating the distribution of the test statistics with a t-distribution (and thus avoids approximating the degrees of freedom).

The degree of smoothing in the STMM is determined by a combination of the spatial dependence parameters and the variance components. At the extreme, if the vertex variance component is equal to zero, then the vertex random effects are equal to zero such that activation is constant across the region. This does occur in a few parcels in our data application, suggesting greater homogeneity in these regions. In our simulations, the STMM had much greater power to detect activation than the MUMM when the vertex-random effects equaled zero. Thus when smoothing is "complete," power appears to

increase, which is consistent with previous studies using a subject-vertex random effect with no vertex random effect (Derado et al., 2010; Bernal-Rusiel et al., 2013). As the vertex variance component increases, activation in the region becomes more heterogeneous, power appears to decrease, and the amount of smoothing is a balance between the degree of spatial dependence and the size of the vertex variance component, as can be seen in (3.25). Overall, the power to detect activation at an individual vertex appears to be most affected by the size of the vertex variance component. For a fixed non-zero vertex variance component, the power tended to decrease as the spatial dependence increased (Table 3.4); this result parallels the effect of serial dependence on the marginal variance and inference in time-series models.

Our model could be applied to irregularly spaced data including subject-specific distances, whereas the spatially autoregressive models common in Bayesian analyses would not be able to exploit this additional information. In our random effects approach, the vertices are viewed as a random sample from an infinite population of vertices. In our context, the potentially infinite population corresponds to locating points anywhere on the cortex. Then our estimates of the random effects, $\hat{\boldsymbol{u}}$, are a conditional mean given the specific locations in the data. This is the typical setting in geostatistics and kriging in particular, where we define the covariance structure for any two points in the continuous spatial domain. In contrast, one would not be able to apply the spatially autoregressive models popular in Bayesian approaches on irregular grid locations. Our model extends quite naturally to using subject-specific distances as available from the mid-thickness cortical sheet, which would be the best estimate of the actual locations of vertices within an individual's brain. The subject-specific approach could more accurately capture spatial dependencies. We do not explore this potential benefit in our application, but instead we use the spherical template (equivalent for all subjects), which is computationally less burdensome.

There are a few approaches to explore that could potentially reduce the MSE of our estimators. Perhaps the easiest is to use $(\mathbf{X}_n^{*\prime}\widehat{\mathbf{\Psi}}_{nv}^{-1}\mathbf{X}_n^*)^{-1}\mathbf{X}_n^{*\prime}\widehat{\mathbf{\Psi}}_{nv}^{-1}$ rather than $(\mathbf{X}_n^{*\prime}\mathbf{X}_n^*)^{-1}\mathbf{X}_n^{*\prime}$ in the definition of $\mathbf{K}_{nv}$ and $\boldsymbol{d}_{nv}$ from (3.11). If we knew $\mathbf{\Psi}_{nv}$ exactly, then a GLS-like transformation would be more efficient. Since we do not know the true covariance, this alternative dimension-reduction technique would introduce a random matrix into the transformation, which would make it difficult to calculate method-of-moments estimators. However, we could simply use the estimators derived from the deterministic projection but substitute $(\mathbf{X}_n^{*\prime}\widehat{\mathbf{\Psi}}_{nv}^{-1}\mathbf{X}_n^*)^{-1}\mathbf{X}_n^{*\prime}\widehat{\mathbf{\Psi}}_{nv}^{-1}$ for $(\mathbf{X}_n^{*\prime}\mathbf{X}_n^*)^{-1}\mathbf{X}_n^{*\prime}$, although there do not appear to be theoretical guarantees that such an approach would yield improved estimators. Another approach that could be attempted is REML on the dimension-reduced data. The computational infeasibility of REML approaches is likely to endure for the foreseeable future, since the likelihood generated from $\mathbf{Y}$ involves an $NVT \times NVT$ matrix (10,806,560 for our largest parcel, and it is not sparse). However, we could derive a Newton-Rhapson or Fisher-scoring algorithm to attempt to fit the likelihood containing the $NVQ \times NVQ$ covariance matrix formed from the reduced-dimension data. Using the estimators presented in this paper as good starting values, such an approach could potentially require relatively few iterations and be computationally feasible, although it is unclear whether it would lead to estimators with substantively lower MSE.

APPENDIX A

APPENDIX TO CHAPTER 1

A.1 Simulation studies

A.1.1 The Infomax algorithm

We are not aware of functions or packages in R that implement the Infomax algorithm (Bell and Sejnowski 1995). We offer an alternative to Matlab code (`http://cnl.salk.edu/~tewon/ICA/code.html`), but with a few modifications that decrease computation time. First, we use the full data (the so-called offline algorithm) in each iteration rather than an online algorithm with batches. Secondly, we use an adaptive method to choose the step size (based upon Bernaards and Jennrich 2005), which speeds up convergence. We also omitted the bias term (intercept) included in the original formulation because we centered our data. R code implementing the Infomax algorithm and example simulations are available by request.

A.1.2 The ProDenICA algorithm

We made small modifications in the simulated data analysis in order to use the R-package ProDenICA. When the IC density was heavy-tailed (e.g., t-distribution with $df = 3$ or $df = 5$), the algorithm sometimes failed in the density estimation step. These issues were resolved by removing one or more of the most extreme outliers.

It should be noted that the 'restarts' option in the `ProDenICA()` function evaluates the objective function at N random matrices, determines the matrix with the highest

negentropy, and then initiates the ProDenICA algorithm with this single matrix. We found that `ProDenICA()` should instead be initiated using multiple random matrices because a single initial value may have a relatively high initial negentropy but be in a basin with a local maximum.

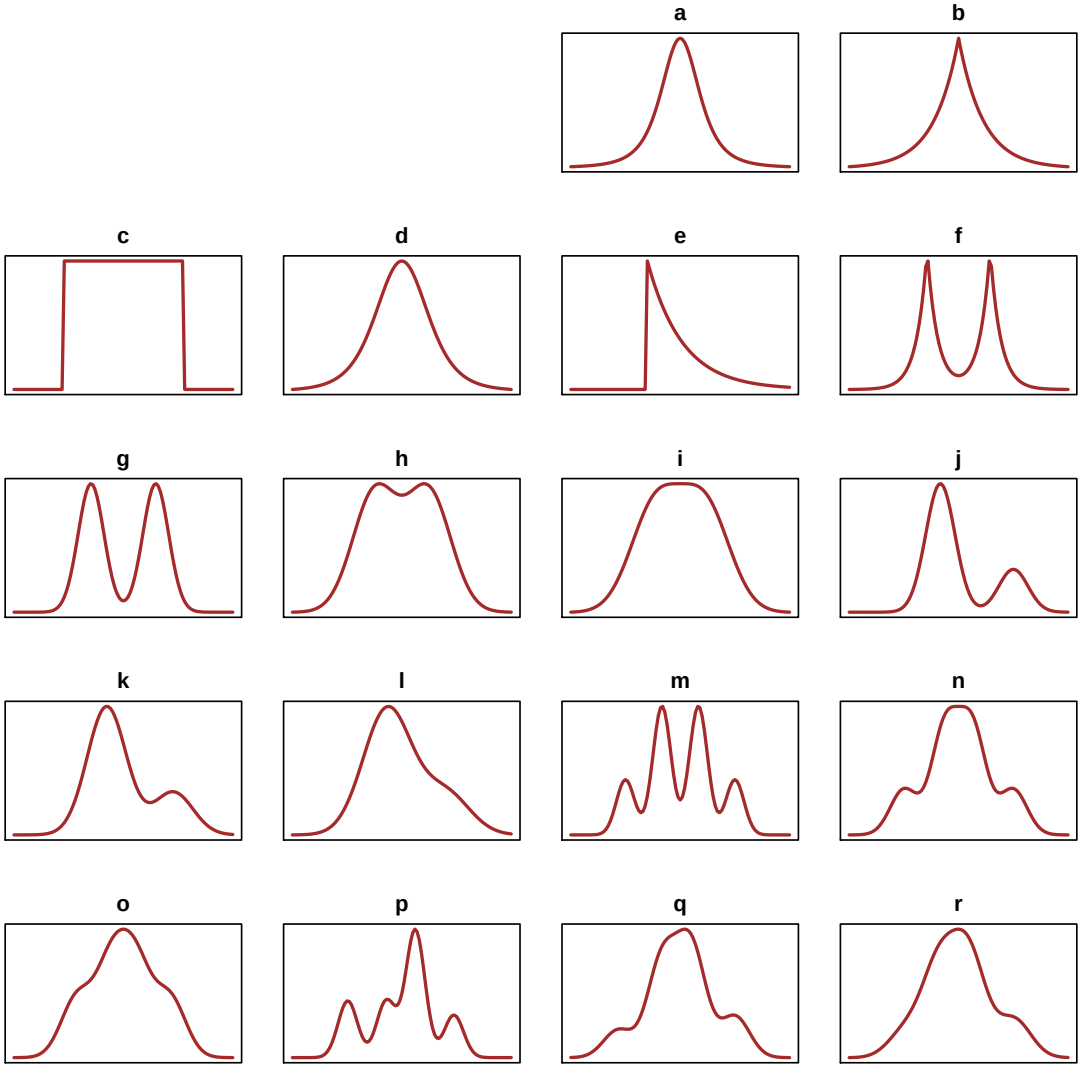
Another issue that arose is that `ProDenICA()` produced an error when using the whitening option with $Q < T_r$. This issue was resolved by supplying `ProDenICA()` with an initial unmixing matrix (rather than relying upon the default).

Lastly, we found that when using the log cosh nonlinearity (`ProDenICA()` provides a function that replicates `fastICA()`), the negentropy measure was not correct; it calculated the mean of $\frac{1}{\alpha}\log\cosh(\alpha s)$. It should instead apply the formula in (1.6).

A.1.3 Simulated data

We simulated the mixing matrix $\mathbf{A}$ using the `mixmat()` function from the R package ProDenICA (Hastie and Tibshirani 2010), which ensures the condition number is between 1 and 2 by simulating a $Q \times Q$ matrix with iid entries from a standard normal, taking the SVD, then generating random eigenvalues from the uniform(1,2) distribution, and defining $\mathbf{A}$ as the product of the left eigenvector, these new eigenvalues, and the right eigenvector. We conducted 100 simulations with $V = 1,024$ samples for each component. Twenty-five initial values were used for the iterative methods, where initial values were randomly selected from a latin hypercube using the angular (Givens) parameterization, with $\theta_q \in [0, \pi]$ for $q = 1, \ldots, Q(Q-1)/2 - 1$ and $\theta_{Q(Q-1)/2} \in [0, \pi/2]$. Data were simulated from eighteen distributions using `rjordan()` in the ProDenICA package (Hastie and Tibshirani 2010; Figure A.1).

Figure A.1: Distributions used in simulations, which include the t-distribution with df=3, double exponential, uniform, t-distribution with df=5, exponential, a mixture of exponentials, and numerous mixtures of normals. Note that a, b, d, and e are super-Gaussian, while c and f - r are sub-Gaussian.



A.1.4 Notes on the minimum distance measure

We adapt the minimum distance (MD) measure (Ilmonen et al. 2010), which was defined for some estimate $\widehat{\mathbf{W}}_{(i)}$ when the true unmixing matrix, $\mathbf{W}$, is known. We apply the measure to two arbitrary square matrices $\mathbf{B}_{(i)}$ and $\mathbf{B}_{(j)}$. Let $\mathcal{P}$ denote the set of $Q \times Q$ signed permutation matrices and C the set of $Q \times Q$ full-rank diagonal matrices. Then define the set of scaled permutation matrices $\mathcal{K} = \{\mathbf{K} : \mathbf{K} = \mathbf{PC},\ \forall\ \mathbf{P} \in \mathcal{P},\ \mathbf{C} \in C\}$. Then the minimum distance measure between two matrices $\mathbf{B}_{(i)}$ and $\mathbf{B}_{(j)}$ is

$$d_{MD}(\mathbf{B}_{(i)}, \mathbf{B}_{(j)}) = \frac{1}{\sqrt{Q-1}} \inf_{\mathbf{K} \in \mathcal{K}} \| \mathbf{K}\mathbf{B}_{(i)}\mathbf{B}_{(j)}^{-1} - \mathbf{I}_d \|_F$$

where $\| \cdot \|_F$ denotes the Frobenius norm. Code implementing this measure is available in the R package *JADE* (Nordhausen et al. 2011).

A.1.5 Computation times

We conducted our simulations on a cluster of 28 Dell PowerEdge 2650 servers with 8 processors per server, where each processor was 2.66 GHz. We used the R package *snow* (Tierney et al. 2011) to conduct simulations in parallel. Computation times are presented in Table A.1.

A.2 Matching ICs

Our approach to matching ICs follows a modification of the Hungarian (Kuhn-Munkres) algorithm (Tichavsky and Koldovsky 2004), and here we describe the modification in detail. Suppose we want to compare $\widehat{\mathbf{S}}^k_{(i)} \in \mathbb{R}^{V \times Q}$ and $\widehat{\mathbf{S}}^l_{(j)} \in \mathbb{R}^{V \times Q}$, the ith estimate from method k and the jth estimate from method l. Hereafter, we drop the k and l

Table A.1: The 0.025, 0.500, and 0.975 quantiles of computation times (in seconds) based on 100 simulations with 25 initial values per simulation. Quantiles are based on the pooled sample of 2,500 computation times for all methods except for JADE, which is not initialized with multiple starting values and is consequently based on 100 samples.

Q	Quantile	FastICA	Infomax	JADE	ProDenICA
5	0.025	0.01	1.28	0.02	3.43
5	0.500	0.03	3.19	0.02	5.84
5	0.975	1.58	5.95	0.05	30.67
10	0.025	0.04	5.88	0.10	11.70
10	0.500	0.34	11.69	0.17	28.75
10	0.975	2.85	13.05	0.27	267.23
20	0.025	1.11	18.75	2.41	95.66
20	0.500	7.46	25.36	3.98	544.92
20	0.975	27.07	29.02	10.00	2478.45

superscripts to simplify notation, noting that the estimates may or may not be from the same method. Assume that $\widehat{\mathbf{S}}_{(i)}$ is in canonical form, as defined in Section 4.2. We refer to the canonically ordered $\widehat{\mathbf{S}}_{(i)}$ as the template. Let $\widehat{\mathbf{S}}_{(i),q}$ be the qth column of $\widehat{\mathbf{S}}_{(i)}$ and $\widehat{\mathbf{S}}_{(j),r}$ be the rth column from $\widehat{\mathbf{S}}_{(j)}$, and let $\|\cdot\|$ denote the Euclidean norm. Create a $Q \times Q$ distance (cost) matrix $\mathbf{C}$ between the components with elements

$$c_{q,r} = \min\left(\|\widehat{\mathbf{S}}_{(i),q} - \widehat{\mathbf{S}}_{(j),r}\|, \|\widehat{\mathbf{S}}_{(i),q} + \widehat{\mathbf{S}}_{(j),r}\|\right),$$

and define the matrix $\mathbf{B}$ with

$$b_{q,r} = \begin{cases} -1 & \text{if } \min\left(\|\widehat{\mathbf{S}}_{(i),q} - \widehat{\mathbf{S}}_{(j),r}\|, \|\widehat{\mathbf{S}}_{(i),q} + \widehat{\mathbf{S}}_{(j),r}\|\right) = \|\widehat{\mathbf{S}}_{(i),q} + \widehat{\mathbf{S}}_{(j),r}\|, \\ 1 & \text{if } \min\left(\|\widehat{\mathbf{S}}_{(i),q} - \widehat{\mathbf{S}}_{(j),r}\|, \|\widehat{\mathbf{S}}_{(i),q} + \widehat{\mathbf{S}}_{(j),r}\|\right) = \|\widehat{\mathbf{S}}_{(i),q} - \widehat{\mathbf{S}}_{(j),r}\|. \end{cases}$$

Let $\mathbf{S}$ be the set of all permutations of the integers 1 to Q, where for some $\sigma \in \mathbf{S}$, we denote the permutation $\sigma = (\sigma(1), \ldots, \sigma(Q))$. We then use the Hungarian algorithm

(Kuhn 1955) to identify the set such that

$$\sigma^* = \underset{\sigma \in \mathcal{S}}{\operatorname{argmin}} \sum_{q=1}^{Q} c_{q,\sigma(q)}.$$

Then define the signed permutation matrix $\mathbf{P}_1$ with entries $p_{q,a_q} = b_{q,a_q}$ at row q and column a_q, and 0 otherwise. Note that $\mathbf{P}_1$ is equivalent to $\underset{\mathbf{P} \in \mathcal{P}}{\operatorname{argmin}} \parallel \widehat{\mathbf{S}}_{(i)} - \widehat{\mathbf{S}}_{(j)}\mathbf{P}' \parallel_F$.

The method used here to match ICs creates a one-to-one mapping of components. Note that when multiple ICs are being compared, the matching algorithm may be sensitive to the choice of template. In our application, we found that using the estimates from JADE, Infomax, or ProDenICA as the template with one-at-a-time matching resulted in the same ordering as using the FastICA estimate as the template. In situations in which ICs from more than two estimates differ greatly, a method to simultaneously match all ICs could be pursued.

A.3 Group ICA of the ADHD-200 Sample

A.3.1 Resting-state fMRI dataset

Data were selected for analysis from the ADHD-200 Data Sample, which consists of rs-fMRI data from children and adolescents (ages 7-21) from 8 independent sites comprising 491 typically developing subjects and 285 that were diagnosed with ADHD (Table A.2). Subjects were diagnosed with three ADHD subtypes: Inattentive; Hyperactive and Impulsive; and Combined (Hyperactive/Impulsive and Inattentive). However, there were only a total of 11 subjects with ADHD-Inattentive, and half the sites did not have subjects with this diagnosis.

We restricted our analysis to (1) subjects with no recorded history of drug therapy;

Table A.2: Subject diagnosis by site in the ADHD-200 Sample: Typ=Typically Developing; ADHD-C=ADHD-Combined; ADHD-H/Im=ADHD-Hyperactive and Impulsive; ADHD-In=ADHD-Inattentive; WH= Withheld.

Site	Typ	ADHD-C	ADHD-H/Im	ADHD-In	WH
Bradley Hospital/Brown University	0	0	0	0	26
Kennedy Krieger Institute	61	16	1	5	11
NeuroIMAGE Sample	23	18	6	1	25
NYU Child Study Center	99	77	2	44	41
Oregon Health & Science University	42	23	2	12	34
Peking University	116	29	0	49	51
University of Pittsburgh	89	0	0	0	9
Washington University in St. Louis	61	0	0	0	0
Total	491	163	11	111	197

(2) subjects that were right-hand dominant; (3) images with no quality control flags; and (4) subjects that were either ADHD-Combined or ADHD-Inattentive (but not ADHD-Hyperactive and Impulsive). Subjects were classified using either (1) the ADHD Rating Scale IV, (2) Conner's Parent Rating Scale-Revised (Long Version), or (3) Conner's Rating Scale, 3rd edition. Within these scales, there was a small degree of overlap in the intermediate values between subjects diagnosed as typically developing and subjects diagnosed with ADHD, whereas individuals with low values were strictly labeled typically developing and individuals with high values were strictly diagnosed with ADHD. We excluded subjects with scores that we deemed borderline, that is, both control and ADHD subjects that were near the threshold at which ADHD was diagnosed. Specifically, we excluded subjects with ADHD Rating Scale IV values between 36 and 45; Conner's Parent Rating Scale-Revised (Long Version) between 56 and 65; or Conner's Rating Scale between 55 and 66 (Table A.3).

Table A.3: Subjects used in analysis. Typ=Typically Developing; ADHD-C=ADHD-Combined; ADHD-In=ADHD-Inattentive.

Site	Typ	ADHD-C	ADHD-H/Im	ADHD-In	WH
Peking University	86	13	0	19	0
Kennedy Krieger Institute	40	7	0	3	0
NYU Child Study Center	56	16	0	11	0
Oregon Health & Science University	24	8	0	1	0
Total	206	44	0	34	0

Details of the primary image processing pipeline were previously reported (Section 2.1, Eloyan et al. 2012). Processing followed the functional connectome processing scripts on the FCP/INDI site (Mennes et al. 2012). In addition, we aggregated the MNI 152 T1 3 mm template to result in $6 \times 6 \times 6$ mm voxels. We retained the $6 \times 6 \times 6$ mm voxels for which all eight of the voxels in the MNI 152 T1 3 mm template were brain tissue. This resulted in $V = 7,825$ for all subsequent analyses. For subjects in which there were multiple scanning sessions, we only used the first session.

We also used our own whitening function to produce the input data for all algorithms, available in `EvaluatingICA_Rsources.R` upon request. Note that the functions `fastICA()` and `JADE()` automatically whiten data; consequently, we modified their source code to prevent additional whitening.

A.3.2 Differences between algorithms

We compared the unmixing matrices from FastICA, Infomax, JADE, and ProDenICA using the MD measure, the Amari measure, and the Frobenius distance between matched unmixing matrices. To aid in our interpretation of the magnitude of differences between

mixing matrices, we simulated the distribution of these three measures for randomly generated orthogonal matrices using two methods. First, orthogonal matrices were generated with columns equal to the eigenvectors from the spectral decompositions of randomly generated matrices following a Wishart distribution with covariance equal to the identity matrix and V degrees of freedom. Second, we simulated uniformly distributed Givens rotation angles $\theta_i \in [-\pi, \pi]$ for $i = 1, \ldots, Q(Q-1)/2$, and then converted the angles to orthogonal matrices (Figure A.4).

In Table A.5, we present false discovery rate (FDR) adjusted p-values from two-sample Kolmogorov-Smirnov tests for equality in distribution between ICs estimated using the SVD, FastICA, Infomax, and ProDenICA. In multiple hypothesis testing, the FDR is the expected proportion of false positives among the rejected null hypotheses, and controlling the FDR leads to more powerful testing procedures than controlling the family-wise error rate (Benjamini and Hochberg 1995). For each p-value, we calculated an FDR-adjusted p-value, called a *q-value* (Storey 2002): let G denote the total number of tests and $p_{(g)}$ denote the gth order statistic from the set of all G p-values, and define the q-value

$$p^*_{(g)} = \min\left(\frac{G}{g} p_{(g)}, p^*_{(g+1)}, \ldots, p^*_{(G)}, 1\right).$$

In typical applications, $p^*_{(g)}$ is an estimate of the minimum proportion of false positives given that at least one rejection occurs, where the minimum is taken over all rejection regions containing $[0, p_{(g)}]$. Here, we use the FDR-adjusted p-values as a measure of the difference between IC distributions since the test statistics were based on spatially dependent data.

We also present density plots for each IC and each method (Figure A.2). Densities were estimated using a Gaussian kernel. For each component, a bandwidth was determined for the estimate from FastICA, Infomax, JADE, and ProDenICA, respectively,

Table A.4: Distance and measures between unmixing matrices by method for the rs-fMRI study. Here, the SVD mixing matrix is taken to be the identity matrix. MD = Minimum Distance measure. Mean and 1% Wishart denote the mean and 1% quantiles, respectively, of each measure from matrices randomly generated via the SVD of iid Wishart matrices. Mean and 1% unif denote the corresponding statistics for matrices generated from the angular parametrization of orthogonal matrices with angles uniformly distributed in $[-\pi, \pi]$.

Method.1	Method.2	Amari	MD	Frobenius
Mean: Wishart 1	Wishart 2	0.35	0.90	6.31
1%: Wishart 1	Wishart 2	0.31	0.88	5.92
Mean: Unif 1	Unif 2	0.26	0.85	6.32
1%: Unif 1	Unif 2	0.22	0.80	5.76
SVD	fastICA	0.36	0.91	6.30
SVD	Infomax	0.36	0.91	6.33
SVD	JADE	0.35	0.90	6.30
SVD	ProDenICA	0.33	0.89	6.29
FastICA	Infomax	0.01	0.07	0.29
FastICA	JADE	0.06	0.38	1.75
FastICA	ProDenICA	0.06	0.41	1.89
Infomax	JADE	0.06	0.39	1.80
Infomax	ProDenICA	0.06	0.42	1.93
JADE	ProDenICA	0.07	0.41	1.85

using the method of Sheather and Jones (1991), and then these four bandwidths were averaged, and finally the densities were estimated with bandwidth fixed at this average. Thus, for a given component, the densities for each of the methods were estimated using the same bandwidth.

Figure A.2: Density plots of ICs for FastICA, Infomax, JADE, and ProDenICA. Values on the x-axis correspond to the standardized BOLD signal. The sample skewness and kurtosis from the FastICA estimates are included in the plot area.

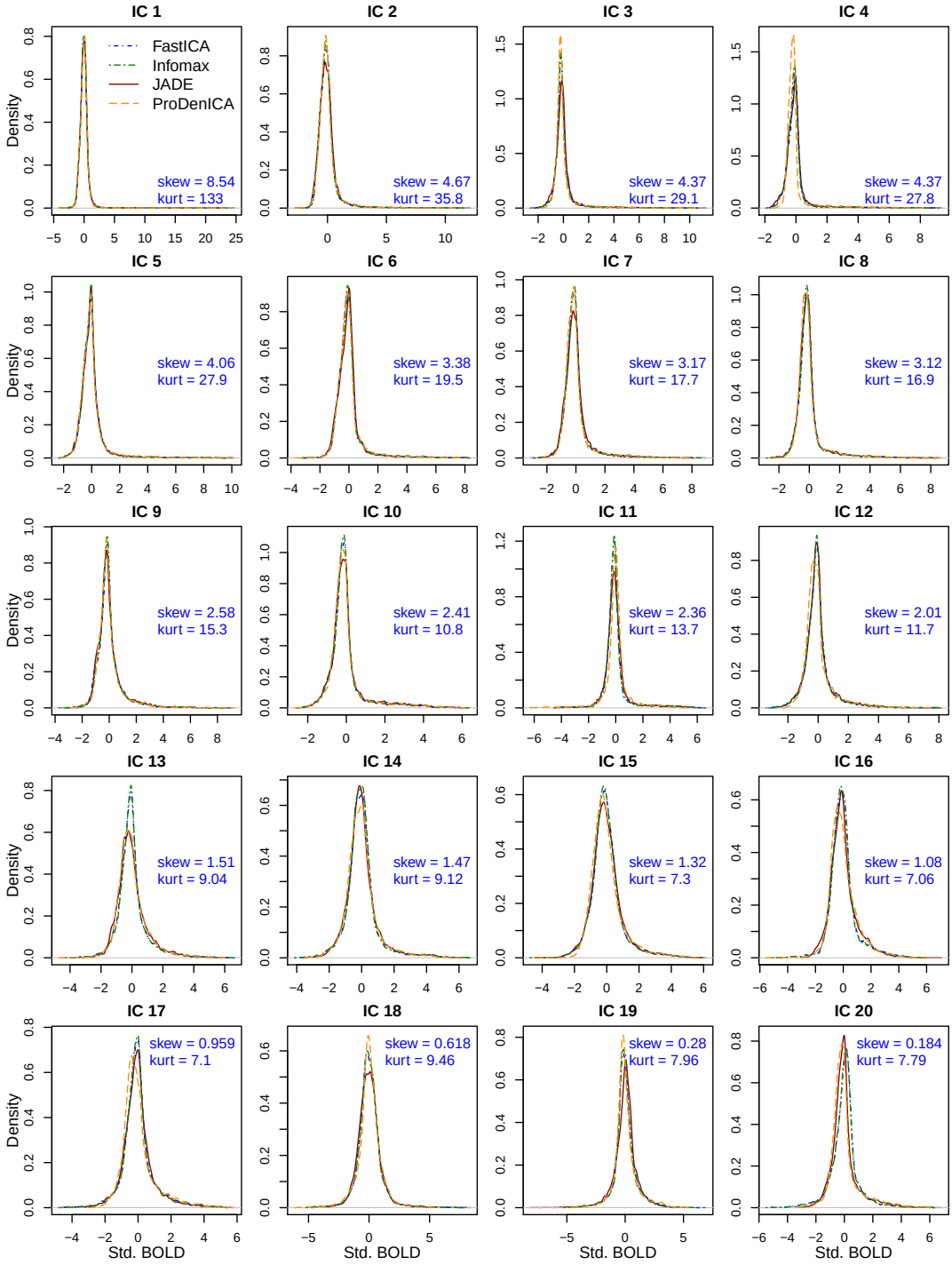


Table A.5: FDR-adjusted p-values from two-sample Kolmogorov-Smirnov statistics. Blank entries indicate FDR-adjusted $p < 0.0001$.

Method1	Method2	IC 1	IC 2	IC 3	IC 4	IC 5	IC 6	IC 7	IC 8	IC 9	IC 10
SVD	FastICA										
SVD	Infomax										
SVD	JADE										
SVD	ProDenICA										
FastICA	Infomax	0.9826	0.2733	0.1277		0.0556	0.5650	0.3543	0.4036	0.1105	0.9481
FastICA	JADE	0.6165	0.0101		0.4788	0.2421	0.0001	0.0004	0.0222	0.0003	0.0129
FastICA	ProDenICA	0.0658	0.0166			0.0451	0.1277	0.0002		0.0053	0.0129
Infomax	JADE	0.4688				0.1574			0.0556	0.0004	0.0024
Infomax	ProDenICA	0.1370	0.2660			0.2354	0.1811	0.0005		0.0254	0.0027
JADE	ProDenICA	0.2807				0.0254			0.0002	0.0265	0.1415

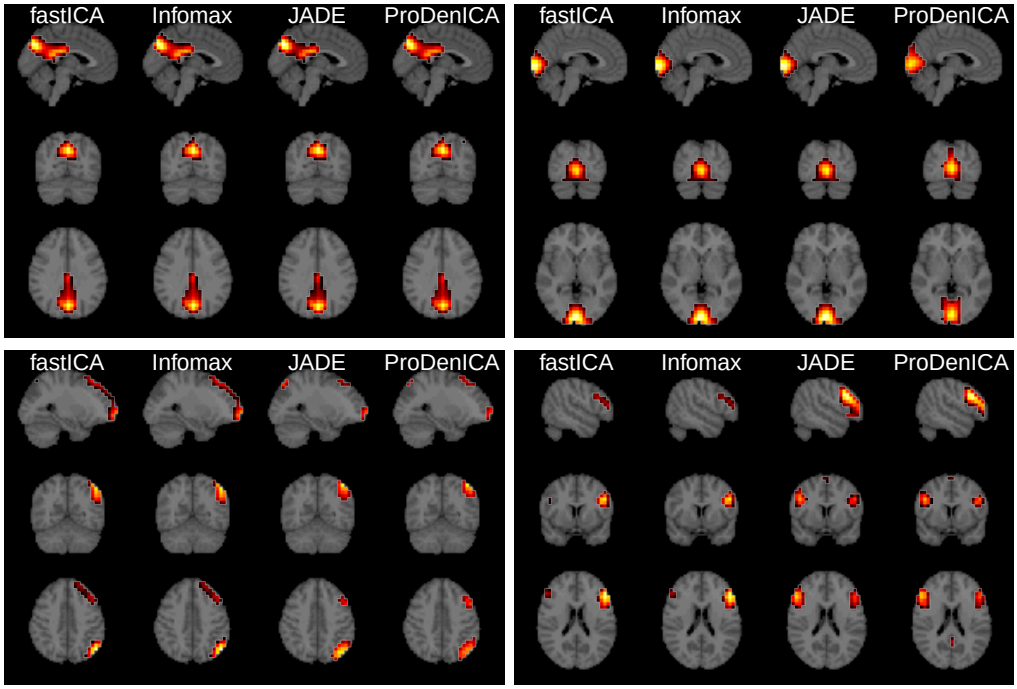
Method1	Method2	IC 11	IC 12	IC 13	IC 14	IC 15	IC 16	IC 17	IC 18	IC 19	IC 20
SVD	FastICA								0.0004		
SVD	Infomax								0.0003		
SVD	JADE								0.0002		
SVD	ProDenICA							0.0018			
FastICA	Infomax	0.0878	0.4890	0.3943	0.3851	0.9826	0.4225	0.7906	0.9826	0.2867	0.4130
FastICA	JADE		0.2136		0.0380	0.1068	0.0101	0.1866	0.0006		
FastICA	ProDenICA										
Infomax	JADE		0.9826		0.0112	0.0433	0.0002	0.0348	0.0002		
Infomax	ProDenICA										
JADE	ProDenICA			0.2867	0.0304						

A.3.3 Selected resting-state networks

Figure A.3 presents images for selected ICs from the group ICA of the ADHD-200 Data Sample. Images were thresholded to retain voxels with values greater than the 97.5% quantile. Slices were chosen to approximately maximize the number of visible activated voxels.

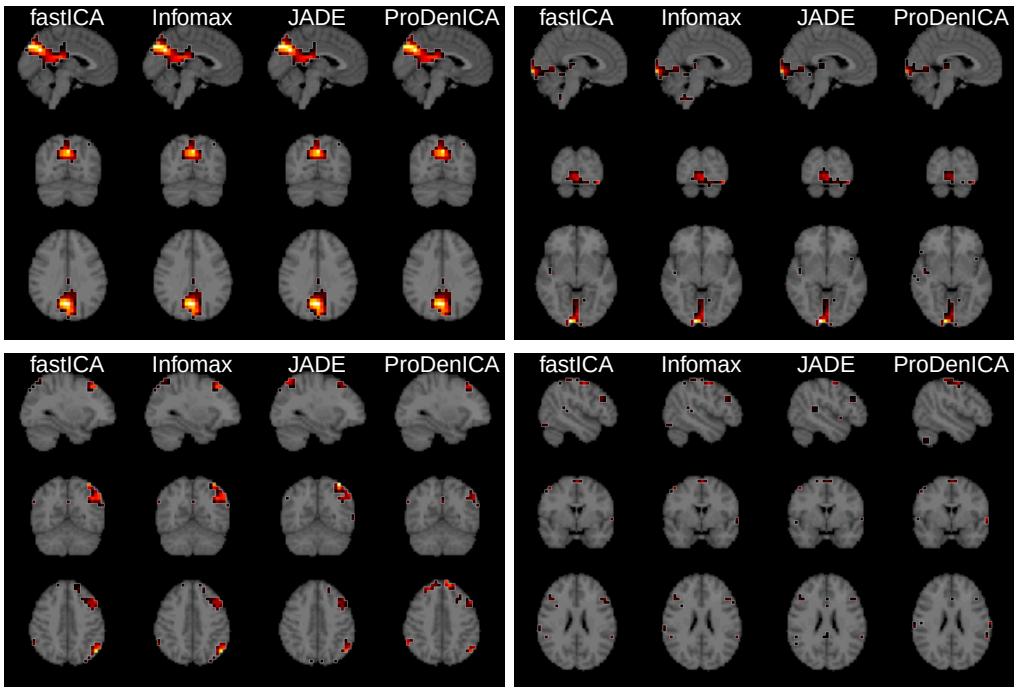
We estimated ICs from a single individual randomly chosen from the ADHD-200 Sample (subject ID 3446674.1.1.pek2). We matched the FastICA estimates from this in-

Figure A.3: Estimated ICs. Clockwise from the top-left: IC 3 (parts of default network), IC 4 (parts of the visual cortex), IC 13 (strong lateralization for FastICA and Infomax but not JADE and ProDenICA), and IC 20 (strong lateralization in all methods).



dividual to the skewness-ordered FastICA estimates of the group ICs, and then matched the ICs from Infomax, JADE, and ProDenICA to these re-ordered FastICA results. Selected ICs are presented in Figure A.4.

Figure A.4: Estimated ICs for a single subject randomly chosen from the ADHD-200 Sample (subject ID 3446674.1.1.pek2). Clockwise from the top-left: IC 3 (parts of default network), IC 4 (medial areas of the visual cortex), IC 13, and IC 20.



APPENDIX B

APPENDIX TO CHAPTER 2

B.1 Using the fixed-point algorithm to fit the LCA model

Here we describe the fixed-point algorithm from Hyvarinen (1999). Our account is equivalent to Hyvarinen (1999) except for our orthogonalization method. Under the constraint that the noise components follow a standard normal distribution, we can ignore rows $\hat{Q} + 1 : T$ in $\widehat{\mathbf{W}}$. For now, we assume the densities of the latent components $f_1, \ldots, f_{\hat{Q}}$, are known. Define the scalar $h_q(x) = \log f_q(x)$, and let $h'(x)$ denote its derivative. Here we use $\mathbf{A}^T$ to denote the transpose of $\mathbf{A}$ to avoid confusion with h'. We can then estimate $\widehat{\mathbf{W}}_\mathbf{S}$:

Algorithm 1: The fastICA algorithm.

Inputs : The whitened $V \times T$ data matrix $\mathbf{Z}$; initial $\mathbf{W}_\mathbf{S}^0$; tolerance ϵ.

Result: Estimates of the latent components, $\widehat{\mathbf{S}} = \mathbf{Z}\widehat{\mathbf{W}}_\mathbf{S}'$.

1. Let $\mathbf{S}^{(0)} = \mathbf{Z}\mathbf{W}_\mathbf{S}^{(0)'}$ and let $n = 0$.

2. For each $q = 1, \ldots, Q$, calculate

$$\mathbf{w}_q^* = \frac{1}{V}\sum_{v=1}^{V}\left\{\mathbf{z}_v h_q'(\mathbf{w}_q^{(n)\,T}\mathbf{z}_v) - h_q''(\mathbf{w}_q^{(n)\,T}\mathbf{z}_v)\mathbf{w}_q^{(n)}\right\}$$

3. Calculate the SVD of $\mathbf{W}_\mathbf{S}^* = \mathbf{U}^*\,\mathbf{D}^*\,\mathbf{V}^{*\,T}$.

4. Let $\mathbf{W}^{(n+1)} = \mathbf{U}^*\,\mathbf{V}^{*\,T}$.

5. If $PMSE(\mathbf{W}_\mathbf{S}^{(n+1)}, \mathbf{W}_\mathbf{S}^{(n)}) < \epsilon$, stop, else increment n and repeat (2)-(3).

B.2 Estimation using Spline-LCA

We adapt the ProDenICA algorithm of Hastie and Tibshirani (2003) to LCA in which we alternate between estimating $\mathbf{W_S}$ for fixed $\hat{f}_1, \ldots, \hat{f}_{\widehat{Q}}$ via the fixed point algorithm and estimating $f_1, \ldots, f_{\widehat{Q}}$ for fixed $\widehat{\mathbf{W}}$ using the "Poisson trick". Our account largely follows the description in Hastie et al. (2009) but with a few departures as noted. Let $\mathbf{Z}$ denote the whitened data. Consider the penalized log likelihood without the noise components:

$$\ell(\mathbf{W_S}, g_1, \ldots, g_Q; \mathbf{z}_1, \ldots, \mathbf{z}_V) =$$
$$\sum_{q=1}^{Q} \left[- \int \phi(x) e^{g_q(x)} dx - \lambda_q \int \{g_q''(x)\}^2 dx + \frac{1}{V} \sum_{v=1}^{V} \left\{ g_q(\mathbf{w}_q^T \mathbf{z}_v) + \log \phi(\mathbf{w}_q^T \mathbf{z}_v) \right\} \right]. \quad \text{(B.1)}$$

Recall that updating $\mathbf{W_S}$ requires the first and second derivatives of the log densities of the latent components, which makes the use of B-splines convenient.

For density estimation, suppose $\mathbf{W_S}$ is given and define $s_{vq} = \mathbf{w}_q' \mathbf{z}_v$. Let $x_1^*, \ldots, x_{L+1}^*$ define a discretization, $[x_1^*, x_2^*), [x_2^*, x_3^*), \ldots, [x_L^*, x_{L+1}^*)$, of the support of the tilt function of the non-Gaussian densities such that $\Delta = x_\ell^* - x_{\ell-1}^*$ for all $\ell = 2, \ldots, L+1$. It suffices to take $x_1^* = \min(s_{11}, \ldots, s_{nd}) - 0.1\hat{\sigma}_z$ and $x_{L+1}^* = \max(s_{11}, \ldots, s_{nd}) + 0.1\hat{\sigma}_z$, where $\hat{\sigma}_z$ denotes the sample standard deviation, which here is equal to one. Next, let $x_\ell = \frac{1}{2}(x_\ell^* + x_{\ell+1}^*)$. For each $q \in \{1, \ldots, \widehat{Q}\}$ and $\ell \in \{1, \ldots, L\}$, define

$$y_{\ell q} = \sum_{v=1}^{V} \mathbb{1}\{s_{vq} \in [x_\ell^*, x_{\ell+1}^*)\}.$$

We approximate (B.1) by discretizing the first integral and estimating the sum over V as

a weighted sum over L. Restricting our attention to a single q, we have

$$-\lambda_q \int \left\{g_q''(x)\right\}^2 dx + \sum_{\ell=1}^{L} \left[\frac{y_{\ell q}}{V}\left\{g_q(x_\ell) + \log \phi(x_\ell)\right\} - \Delta\phi(x_\ell)e^{g_q(x_\ell)}\right].$$

and dividing by Δ, we have

$$\beta_q \int \left\{g_q''(x)\right\}^2 dx + \sum_{\ell=1}^{L} \left[\frac{y_{\ell q}}{V\Delta}\left\{g_q(x_\ell) + \log \phi(x_\ell)\right\} - \phi(x_\ell)e^{g_q(x_\ell)}\right]. \tag{B.2}$$

for some penalty β_q. This is proportional to a Poisson generalized additive model (GAM), where $\frac{y_{\ell q}}{V\Delta}$ is the response and the expected response is equal to $\phi(x_\ell)e^{g_q(x_\ell)}$. This can be fit using the `gam` package in R (Hastie, 2013) where β_q is chosen to result in a user-specified number of (approximate) degrees of freedom. We find that $df = 8$ and $L = 100$ produce fast and accurate density estimates in simulations for a variety of densities when the sample size is equal to 1,000. This method also easily scales to tens of thousands of observations since it is $O(V)$, where the main expense is calculating $y_{\ell q}$.

The algorithm to estimate both $\mathbf{W_S}$ and $f_1, \ldots, f_Q$ is summarized below:

Algorithm 2: Estimating the semiparametric LCA.

Inputs : The whitened $V \times T$ data matrix $\mathbf{Z}$; initial $\mathbf{W}_\mathbf{S}^0$; tolerance ϵ.

Result: Estimates of the latent components, $\widehat{\mathbf{S}}$, and their densities, $\hat{f}_1, \ldots, \hat{f}_Q$.

1. Let $n = 0$ and define $\mathbf{S}^{(n)} = \mathbf{Z}\mathbf{W}_\mathbf{S}^{(n)'}$.
2. Estimate $f_q^{(n+1)}$ for $q = 1, \ldots, Q$.
3. Update $\mathbf{W}_\mathbf{S}^{(n+1)}$ given $f_1^{(n+1)}, \ldots, f_Q^{(n+1)}$ and $\mathbf{S}^{(n)}$ with one-step of the fixed-point algorithm.
4. Let $\mathbf{S}^{(n+1)} = \mathbf{Z}\mathbf{W}_\mathbf{S}^{(n+1)}$.
5. If $PMSE(\mathbf{W}_\mathbf{S}^{(n+1)}, \mathbf{W}_\mathbf{S}^{(n)}) < \epsilon$, stop, else increment n and repeat (2)-(4).

B.3 Additional Background

B.3.1 Projection Pursuit, D-FastICA, and Non-Gaussian Subspace Analysis

Projection pursuit is an exploratory method for finding low-dimensional representations of multivariate data that reveal interesting patterns and structure (Huber, 1985). Let $\mathbf{x}_v$, $v = 1, \ldots, V$ be a data sample with $\mathbf{x}_v \in \mathbb{R}^T$, and assume $\sum_{v=1}^{V} \mathbf{x}_v = \mathbf{0}$, where $\mathbf{0}$ is the vector of T zeros, and $\frac{1}{V}\sum_{v=1}^{V} \mathbf{x}_v^2 = \mathbf{1}$, where $\mathbf{1}$ is a length T vector of ones. Let Q be the number of projection pursuit directions that are estimated. In FastICA in deflation mode (D-FastICA), the projection pursuit index is equivalent to an approximation of negentropy (Hyvarinen, 1999):

$$\hat{\mathbf{w}}_q = \underset{\mathbf{w}\in\mathbb{R}^T}{\operatorname{argmax}} \left\{ \frac{1}{V}\sum_{v=1}^{V} G(\mathbf{w}'\mathbf{x}_v) - \mathrm{E}\, G(n) \right\}^2 \tag{B.3}$$

where $\mathbf{w}$ is orthogonal to $\hat{\mathbf{w}}_1, \ldots, \hat{\mathbf{w}}_{q-1}$ and $||\mathbf{w}|| = 1$ with $||\cdot||$ denoting the L2-norm, G is a non-linear function, and n is a standard normal random variable. A common choice for G is $\log\cosh(x)$, which will be used to estimate projection pursuit directions in our simulations.

NGCA uses multiple projection pursuit indices (Blanchard et al., 2006) or radial basis functions (Kawanabe et al., 2007) to find a non-Gaussian subspace that is assumed to contain the interesting features of data. NGCA can be formulated using a semiparametric likelihood,

$$f_{\mathbf{X}}(\mathbf{x}) = h(\mathbf{W}_{\mathbf{S}}\mathbf{x})\phi_{\mathbf{0},\boldsymbol{\Sigma}}(\mathbf{x}) \tag{B.4}$$

where $\phi_{\mathbf{0},\boldsymbol{\Sigma}}$ is multivariate normal with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$; $\mathbf{W}_{\mathbf{S}}$ is a $Q \times T$ matrix; and $h(\cdot)$ is a function that captures departures from Gaussianity under the constraint that

$f_{\mathbf{X}}(\mathbf{x})$ is a density. The NGCA model does not assume independent factors, and we do not consider it in our simulations, although we will show that one of our proposed methods can be written in the form of (B.4).

The density in the Spline-LCA model is a special case of (B.4) from NGCA but with the additional assumption of independence.

Proposition 2. *Let* $\mathbf{X}$ *be a random variable from the LCA model where the LCs have tilted Gaussian densities. Then the density of* $\mathbf{X}$ *is*

$$f_{\mathbf{X}}(\mathbf{x}) = \phi_{\mathbf{0},\mathbf{\Sigma}}(\mathbf{x}) \prod_{q=1}^{Q} e^{g_q(\mathbf{w}_q'\mathbf{L}\mathbf{x})}$$

where $\phi_{\mathbf{0},\mathbf{\Sigma}}$ *is the mean zero multivariate distribution with covariance* $\mathbf{\Sigma}$.

Proof. Using the tilted Gaussian density, we have

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \det \mathbf{L} \prod_{q=1}^{Q} e^{g_q(\mathbf{w}_q'\mathbf{L}\mathbf{x})} \phi(\mathbf{w}_q'\mathbf{L}\mathbf{x}) \prod_{k=1}^{T-Q} \phi(\mathbf{w}_{k+Q}'\mathbf{L}'\mathbf{x}) \\ &= \left\{ \prod_{q=1}^{Q} e^{g_q(\mathbf{w}_q'\mathbf{L}\mathbf{x})} \right\} (2\pi)^{-T/2} (\det \mathbf{L}) \exp\left\{ -\frac{1}{2} \sum_{k=1}^{T} \mathbf{x}'\mathbf{L}'\mathbf{w}_k\mathbf{w}_k'\mathbf{L}\mathbf{x} \right\} \\ &= (\det \mathbf{\Sigma})^{-1/2} (2\pi)^{-T/2} \exp\left\{ -\frac{1}{2}\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x} \right\} \prod_{q=1}^{Q} e^{g_q(\mathbf{w}_q'\mathbf{L}\mathbf{x})}. \end{aligned}$$

□

Writing the likelihood in this way makes clear that we are using the Gaussian density to model the covariance between components and we are using the tilt functions to model deviations from the Gaussian model.

B.3.2 Noise-free ICA, PCA-Infomax, and PCA-ProDenICA

In the noise-free ICA model, the number of components is equal to the dimension of the data. Now let $\mathbf{x}_v$, $v = 1, \ldots, V$ be an iid sample. Let $M_{\mathbf{S}}$ be an invertible $T \times T$ matrix, which is called the mixing matrix. The ICA model is,

$$\mathbf{x}_v = \mathbf{M}_{\mathbf{S}} \boldsymbol{s}_v \tag{B.5}$$

where $\boldsymbol{s}_v = (s_{1v}, \ldots, s_{Tv})'$ and the elements of $\boldsymbol{s}_v$ are mutually independent, non-degenerate random variables with at most one component having a Gaussian distribution. Additionally, it is assumed that $\mathrm{E}\, \boldsymbol{s}_v = \mathbf{0}$ and $\mathrm{E}\, \boldsymbol{s}_v^2 = \mathbf{1}$. Under these assumptions, the model is identifiable up to signed permutations of the columns of $\mathbf{M}$ and corresponding rows of $\boldsymbol{s}_v$. Then $\mathbf{S}_q = (s_{q1}, \ldots, s_{qV})'$ is the qth IC. estimated non-parametrically (e.g., Hyvärinen et al. 2001; Samworth and Yuan 2012).

Infomax is a popular noise-free ICA model that can be derived as a maximum likelihood estimator for latent components that have a logistic distribution (Cardoso, 1997; Bell and Sejnowski, 1995).

ProDenICA is a semi-parametric ICA model that estimates the density of the components using cubic B-splines (Hastie and Tibshirani, 2003).

Let $\mathbf{X}$ be the $V \times T$ data matrix. As noted in the introduction, $\mathbf{X}$ is usually dimension-reduced using PCA. Then noise-free ICA is applied to the first Q principal components scaled to have unit variance, which is equivalent to the first Q left singular vectors multiplied by $\sqrt{V}$.

B.3.3 Noisy ICA and IFA

In the noisy ICA model, Q ICs are corrupted by rank-T Gaussian noise, where $Q \leq T$ (Hyvärinen et al., 2001),

$$\mathbf{x}_v = \mathbf{M}_\mathbf{S} \boldsymbol{s}_v + \boldsymbol{\epsilon}_v \tag{B.6}$$

with $\mathbf{x}_v \in \mathbb{R}^T$, $\mathbf{M}$ is $T \times Q$ with $Q \leq T$, $\boldsymbol{\epsilon}_v$ is mean-zero multivariate normal with covariance matrix $\boldsymbol{\Psi}$, and $\boldsymbol{\epsilon}_v$ is independent of $\boldsymbol{s}_v$.

In IFA, (B.6) is estimated under the assumption that the densities of the ICs are Gaussian mixtures (Attias, 1999). In its original formulation, $\boldsymbol{\Psi}$ was an arbitrary positive definite matrix, the IC densities had K_q classes, and the variance of each IC was standardized to unity after each iteration. In our presentation and estimation, we will assume that the covariance of the noise is $\sigma^2\mathbf{I}$ and IC densities are mixtures of two Gaussians, which has been assumed elsewhere (e.g., Guo and Tang 2013; Beckmann and Smith 2004), and enforce the constraint that the IC densities are mean zero with unit variance. Let π_{q1} be the probability that an observation of the qth IC comes from the first class, where the first class has a normal distribution with mean μ_{q1} and variance ν_{q1}. Then the probability, mean, and variance for the second class are $\pi_{q2} = 1 - \pi_{q1}$, $\mu_{q2} = -\frac{\pi_{q1}\mu_{q1}}{\pi_{q2}}$, and $\nu_{q2} = \frac{1-\pi_{q1}\nu_{q1}-\pi_{q1}\mu_{q1}^2}{\pi_{q2}} - \mu_{q2}^2$, respectively. Then the joint density of $\mathbf{x}_v$ can be written

$$f_\mathbf{X}(\mathbf{x}; \mathbf{M}) = \prod_{t=1}^{T} \int \phi_{0,\sigma^2}\left(\mathbf{x}_t - \mathbf{m}_t'\boldsymbol{s}\right) f_\mathbf{S}(\boldsymbol{s}) d\boldsymbol{s}, \tag{B.7}$$

where ϕ_{0,σ^2} is a normal density with mean zero and variance σ^2 and

$$f_\mathbf{S}(\boldsymbol{s}) = \prod_{q=1}^{Q} \left\{\pi_{q1}\phi_{\mu_{q1},\nu_{q1}}(s_q) + \pi_{q2}\phi_{\mu_{q2},\nu_{q2}}(s_q)\right\}.$$

Analytic integration across $\boldsymbol{s}$ is possible. Let k_q be equal one if s_q is in the first class and zero otherwise. Let $\mathcal{K}$ be the set of all possible states for the Q components composed

from the Cartesian product Q-times of the singletons $\{\{0\},\{1\}\}$. Let $\mathbf{k}_j = \{k_1, \ldots, k_Q\}$ denote an element of $\mathcal{K}$, where $j \in \{1, \ldots, 2^Q\}$. Let $\boldsymbol{\mu}(\mathbf{k}_j)$ and $\boldsymbol{\nu}(\mathbf{k}_j)$ denote the conditional means of $\boldsymbol{s}$ given the states $\mathbf{k}_j$. Now define

$$\boldsymbol{\Sigma}(\mathbf{k}_j) = \mathbf{M}\ \mathrm{diag}\{\boldsymbol{\nu}(\mathbf{k}_j)\}\ \mathbf{M}' + \sigma^2\mathbf{I}$$

and

$$\boldsymbol{\mu}^*(\mathbf{k}_j) = \mathbf{M}\boldsymbol{\mu}\ (\mathbf{k}_j).$$

Then the density is

$$f_{\mathbf{X}}(\mathbf{x};\mathbf{M}) = \sum_{\mathbf{k}_j \in \mathcal{K}} \Phi\{\mathbf{x};\boldsymbol{\mu}^*(\mathbf{k}_j),\boldsymbol{\Sigma}(\mathbf{k}_j)\}(\mathbf{x}) \prod_{q=1}^{Q} \pi_{q1}^{k_q}\pi_{q2}^{1-k_q} \tag{B.8}$$

with $\Phi\{\mathbf{x};\boldsymbol{\mu}^*(\mathbf{k}_j),\boldsymbol{\Sigma}(\boldsymbol{k}_j)\}(\mathbf{x})$ multivariate normal with mean $\boldsymbol{\mu}^*(\mathbf{k}_j)$ and variance $\boldsymbol{\Sigma}(\boldsymbol{k}_j)\}(\mathbf{x})$. Then a likelihood can be constructed from (B.8), and given some $\widehat{\mathbf{M}}$, the ICs can be estimated from their conditional means. Alternatively, maximum a posteriori estimates of the ICs could be obtained, though we pursue the former here.

B.4 Supplementary materials for simulations examining distributional and noise-structure assumptions

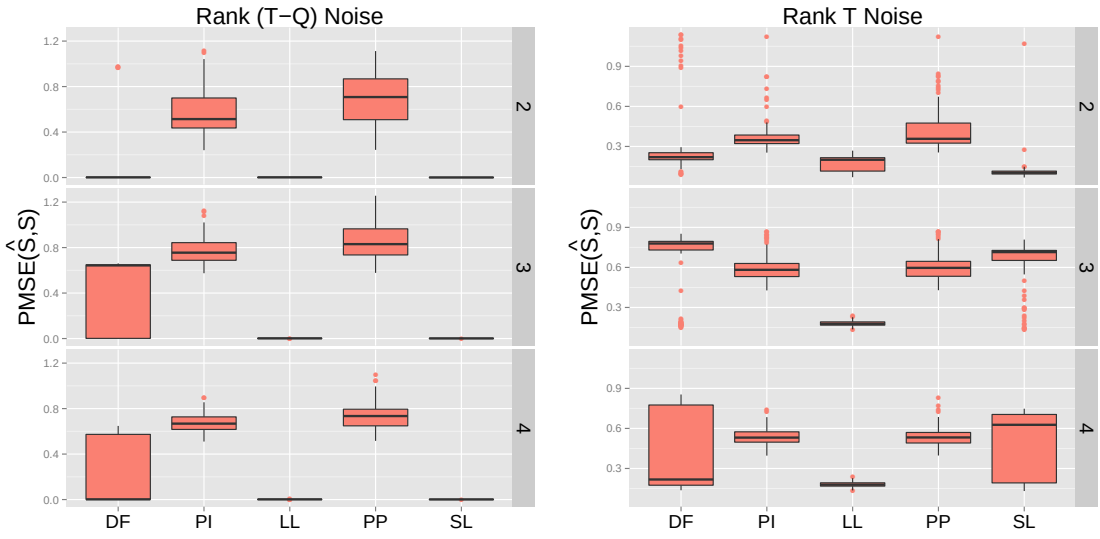
We fit D-FastICA using the 'deflation' option in the fastICA R package (Marchini et al., 2010). However, this popular function does not include an option to use projection pursuit for dimension reduction. If one specifies some $Q < T$ number of components, PCA is performed prior to the ICA. Consequently, one must estimate all T directions and then subset to the first two.

We fit the IFA model with two-class mixtures of normals by maximizing the log likelihood using a numerical optimizer. This contrasts with methods using approximating EM algorithms, as described in the introduction. Our implementation is not scalable

to large Q or T (nor is the exact EM algorithm) but suffices for the simulation experiments. For IFA, one must specify initial values for the unmixing matrix, the variance of the isotropic noise, and the parameters of the Gaussian mixtures. We had four strategies to find the argmax as detailed here. In our function, we constrain the latent component distributions to have zero expectation and unit norm, and as a result, the number of parameters to estimate for each latent component distribution is three. First, we estimated the parameters of the model proposed in Beckmann and Smith (2004) (BS-PICA) and used this solution to initialize the IFA. We then estimated the model from six additional random matrices but with density parameters initialized from the BS-PICA solution. Secondly, when the IFA model was true, we initialized it from the true mixing matrix and true density parameters and also from six additional random matrices with density parameters initialized from their true values. When the IFA model was not true, we initialized it from the true mixing matrix but with the density parameters initialized from their BS-PICA estimates and an additional six random matrices. Thirdly, we initialized the algorithm from seven random matrices but with initial Gaussian mixture densities defined by the parameters (0.7, 0.7, −0.5, −0.5, 0.5, 0.5) (super-Gaussian distribution) for $\pi_{11}, \pi_{21}, \mu_{11}, \mu_{21}, \nu_{11}, \nu_{21}$ and $\sigma^2 = 1$. Finally, we initialized the algorithm from seven random matrices but with initial Gaussian mixture densities defined by the parameters (0.3, 0.3, −1, −1, 0.5, 0.5) (sub-Gaussian distribution) with $\sigma^2 = 1$.

A mixing matrix was generated by first simulating a 5×5 matrix with standard normal entries, taking the singular value decomposition (SVD), then creating a diagonal matrix with five singular values from a uniform(1,10) distribution, followed by multiplying the left singular vectors from the SVD, the diagonal matrix, and the right singular vectors. For the noisy ICA model, we generated a random mixing matrix in the same manner, then retained the first two columns.

Figure B.1: Boxplots of $PMSE$ for estimated columns of $\mathbf{S}$ from simulations of spatial networks with temporal dependence and $Q = 3$. 'DF' = D-FastICA; 'PI' = PCA-Infomax; 'LL'= Logis-LCA; 'PP' = PCA-ProDenICA; 'S-L' = Spline-LCA.



B.5 Supplementary figures for the spatio-temporal network simulations

B.6 Supplementary materials for the fMRI analysis

Whole-brain data were acquired from two sessions with 274 volumes each using gradient-echo EPI with an eight-band multifactor approach and 2 x 2 x 2 mm voxels (repetition time (TR) = 720 ms; echo time (TE) = 33.1 ms; flip angle=52°; field of view = 208 x 180 mm (readout x phase-encoding); acquisition matrix = 104 x 90; slice thickness = 2.0 mm). Only the first session was used in our analyses. The HCP project (Principal Investigators: Bruce Rosen, M.D., Ph.D., Martinos Center at Massachusetts General Hospital; Arthur W. Toga, Ph.D., University of California, Los Angeles, Van J. Weeden, MD, Martinos Center at Massachusetts General Hospital) is supported by

the National Institute of Dental and Craniofacial Research (NIDCR), the National Institute of Mental Health (NIMH), and the National Institute of Neurological Disorders and Stroke (NINDS). Collectively, the HCP is the result of efforts of co-investigators from the University of California, Los Angeles, Martinos Center for Biomedical Imaging at Massachusetts General Hospital (MGH), Washington University, and the University of Minnesota.

After vectorization, the voxels for each matrix for each subject were standardized across time to have mean zero and unit variance. Analyses were initially conducted by concatenating sessions one and two, but subsequent inspection suggested that patterns of network loadings differed greatly between the two sessions. Consequently, only the first session was included. Inspection also revealed that the first two TRs contained BOLD signals that were much higher than other time points, suggesting inadequate equilibration time. Consequently, we removed the first two TRs.

Following Risk et al. (2014), we assessed the reliability of individual components by matching components from all other initializations to the components corresponding to the argmax using the modified Hungarian algorithm. We then created dissimilarity matrices for each component based on the MSE and visualized basins of attraction using multidimensional scaling. Generally, there were at least two basins of attraction corresponding to initializations from the principal subspace and initializations from the entire column space (Supplemental Figure B.2). Components one, two, and nine were relatively robust to initialization and contained only one (main) basin of attraction. Note that in our results, we examined components one and two.

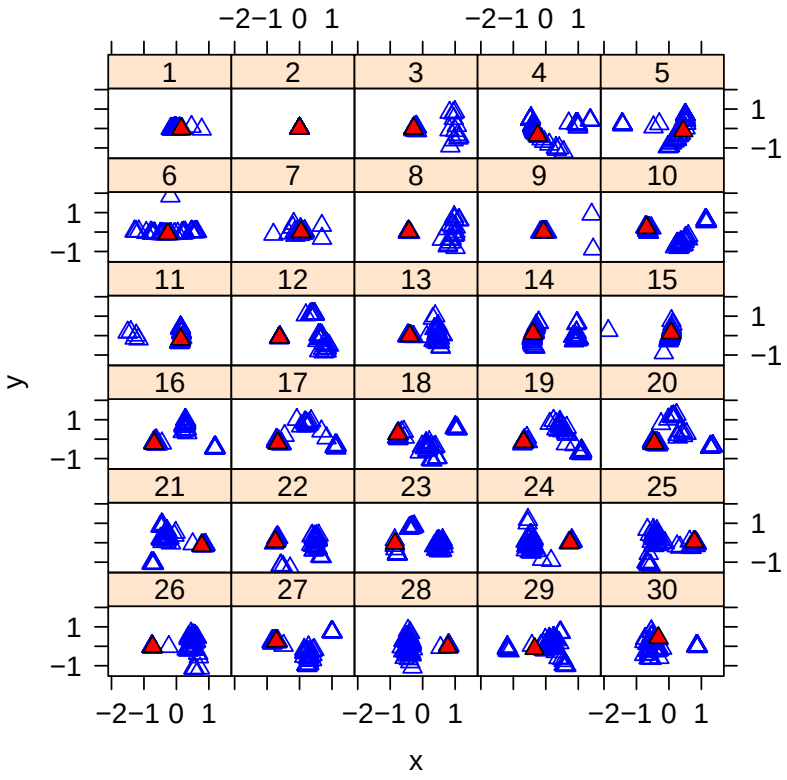


Figure B.2: Multidimensional scaling of $\|\widehat{\mathbf{S}}_j^{(k)} - \widehat{\mathbf{S}}_j^{(l)}\|_2$ for components $j = 1, \ldots, 30$ and initializations $k \neq l \in \{1, \ldots, 30\}$. The coordinates corresponding to the initialization with the highest likelihood are depicted by solid red triangles.

APPENDIX C

APPENDIX TO CHAPTER 3

C.1 Summary of matrix operations and notation

Our notation generally follows Appendix M in Searle et al. (2009).

$\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product.

$\oplus_{n=1}^{N} \mathbf{A}_n$ denotes the direct sum:

$$\oplus_{n=1}^{N} \mathbf{A}_n = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} & \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & & & \mathbf{A}_N \end{bmatrix}.$$

The following notation indicates we are stacking matrices:

$$\{_c\ \mathbf{A}_n\}_{n=1}^{N} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_N \end{bmatrix},$$

where the c denotes that we are forming a column vector in the block-representation. Similarly, we denote concatenation:

$$\{_r\ \mathbf{A}_n\}_{n=1}^{N} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \dots & \mathbf{A}_N \end{bmatrix}.$$

Table C.1: Description of notation. Notation is listed alphabetically with Greek letters alphabetized by their English phonetic spelling (which corresponds to the names used in LaTeX). A notation that is only used once is not included because the definition immediately follows its use.

Notation	Description
a_{nvt}	Error of the nth subject at the vth vertex and tth timepoint.
$\boldsymbol{a}_{nv}$	$[a_{nv1}, \ldots, a_{nvT}]'$
$\boldsymbol{a}_{n}$	$[\boldsymbol{a}'_{n1}, \ldots, \boldsymbol{a}'_{nV}]'$
$\boldsymbol{a}$	$[\boldsymbol{a}'_{1}, \ldots, \boldsymbol{a}'_{N}]'$
b_{nv}	The true value of the vertex-subject interaction random effect for the nth subject at the vth voxel.
$\boldsymbol{b}_{nv}$	$[b_{nv1}, \ldots, b_{nvQ}]'$
$\boldsymbol{b}_{n}$	$[\boldsymbol{b}'_{n1}, \ldots, \boldsymbol{b}'_{nV}]'$
$\boldsymbol{b}_{n}^{q}$	$[b_{n1q}, \ldots, b_{nVq}]'$
$\boldsymbol{b}$	$[\boldsymbol{b}'_{1}, \ldots, \boldsymbol{b}'_{V}]'$
$\mathbf{B}$	$\text{diag}(\sigma^2_{b_1}, \ldots, \sigma^2_{b_Q})$
B	Back-shift operator.
β_{vq}	Fixed effect in the MUMM at the vth vertex for the qth task.
β_{q}	Fixed effect in the STMM.
$\boldsymbol{\beta}_{v}$	$\beta_{v1}, \ldots, \beta_{vQ}$ in the MUMM.
$\boldsymbol{\beta}$	$[\beta_1, \ldots, \beta_Q]'$ in the STMM.
$\mathbf{c}$	Contrast vector for t-statistic
$\mathbf{C}_N$	Centering matrix: $\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}'_N$
d_{nvq}	Projection of $\mathbf{y}_{nv}$ for the qth task in the nth subject at the vth vertex; used in STMM.
$\boldsymbol{d}_{nv}$	$[d_{nv1}, \ldots, d_{nvQ}]'$
$\boldsymbol{d}_{n}$	$[\boldsymbol{d}'_{n1}, \ldots, \boldsymbol{d}'_{nV}]'$
$\boldsymbol{d}$	$[\boldsymbol{d}'_{1}, \ldots, \boldsymbol{d}'_{N}]'$
$\bar{\boldsymbol{d}}_{n\cdot}$	$\boldsymbol{d}_{nv}$ averaged across vertices. Vector in $\mathbb{R}^Q$.
$\bar{\boldsymbol{d}}_{\cdot v}$	$\boldsymbol{d}_{nv}$ averaged across subjects. Vector in $\mathbb{R}^Q$.
$\bar{\boldsymbol{d}}_{\cdot\cdot}$	$\boldsymbol{d}_{nv}$ averaged across subjects and vertices. Vector in $\mathbb{R}^Q$.
$\mathbf{D}_l$	An upper triangular matrix such that $(\mathbf{D}_l)_{ij} = 1$ for $j = i + l$ with $i = l + 1, \ldots, T - l$, and zero elsewhere.
$\delta(a, b)$	Covariogram evaluated for arbitrary random variables a and b.
e_{nvq}	The random plus fixed effect in the hierarchical formulation of the mixed models.
$\boldsymbol{e}_{nv}$	$[e_n v1, \ldots, e_{nvQ}]'$.

Continued on next page

Table C.1 – continued from previous page

Notation	Description
ϵ_{nvt}	Innovation error of the AR process.
η_0	Time-delay parameter of the canonical HRF.
$\mathbf{F}$	$NVQ \times NVQ$ sparse matrix of components of the covariance of the STMM that are independent across subjects.
g	Equivalent to g_q when $Q = 1$.
g_q	$\sum_{v=1}^{V} \sum_{v=1}^{V} \mathbf{\Gamma}_q$.
$\mathbf{G}$	$\text{diag}(g_1, \ldots, g_Q)$, where g_q is the sum of all elements of $\mathbf{\Gamma}_q$.
γ_{nvm}	The fixed effect from the mth nuisance term for the nth subject at the vth vertex
$\boldsymbol{\gamma}_{nv}$	$[\gamma_{nv1}, \ldots, \gamma_{nvM}]'$
$\mathbf{\Gamma}_q$	Spatial correlation matrix for vertex random effects from the qth task.
$\mathbf{\Gamma}$	The $VQ \times VQ$ covariance matrix of $\boldsymbol{u}$
$\mathbf{H}_n$	The hat matrix: $\mathbf{X}_n^*(\mathbf{X}_n^{*\prime}\mathbf{X}_n^*)^{-1}\mathbf{X}_n^{*\prime}$.
$\mathbf{I}_N$	$N \times N$ covariance matrix.
$\bar{\mathbf{J}}_N$	$N \times N$ matrix with all entries equal to $\frac{1}{N}$.
$\mathbf{J}_N$	$N \times N$ matrix of ones.
$\boldsymbol{k}_{nv}$	A vector of length Q equal to the transformed error in the STMM: $\mathbf{K}_n'\boldsymbol{a}_{nv}$.
$\mathbf{K}_n$	The first Q columns of $\mathbf{K}_n^*$.
$\mathbf{K}_n^{*\prime}$	The cap matrix: $(\mathbf{X}_n^{*\prime}\mathbf{X}_n^*)^{-1}\mathbf{X}_n^{*\prime}$.
l	Lag number.
l_0	Nugget effect in the exponential covariogram.
λ_0	Bias parameter in our modified exponential covariogram.
λ_1	Variance parameter in exponential covariogram.
m	Index for nuisance covariate.
M	Total number of nuisance covariates.
n	Index for subject.
N	Total number of subjects.
$\mathbf{\Omega}_q$	Spatial correlation matrix for the subject-vertex interaction random effects from the qth task.
$\mathbf{\Omega}$	The $VQ \times VQ$ covariance matrix of $\boldsymbol{b}_n$.
p	Order of the AR model.
ϕ_{nvp}	The AR coefficient for the pth lag for the nth subject at the vth vertex.
$\psi_{nv}(l)$	Autocorrelation at lag l.
$\mathbf{\Psi}_{nv}$	Correlation matrix of the AR errors for the nth subject at the vth vertex.
q	Index for covariate of interest.
Q	Total number of covariates of interest.
r	Index for region.

Continued on next page

Table C.1 – continued from previous page

Notation	Description
R	Total number of regions.
$\hat{\boldsymbol{r}}_n$	First level residuals: $(\mathbf{I}_T - \mathbf{H}_n)\mathbf{Y}_n$
$\rho_{nv}(l)$	Autocovariance at lag l for the nth subject and vth vertex.
s_{nq}	Subject-specific random slope for the q task and nth subject.
$\boldsymbol{s}_n$	$[s_{n1}, \ldots, s_{nQ}]'$.
$\boldsymbol{s}$	$[\boldsymbol{s}_1', \ldots, \boldsymbol{s}_N']'$
$\mathbf{S}$	$\text{diag}(\sigma_{s_1}, \ldots, \sigma^2_{s_Q})$
$\sigma^2_{b_q}$	Subject-vertex interaction interaction random effect variance for the qth task.
$\sigma^2_{s_q}$	Subject random effect variance for the qth task.
$\sigma^2_{u_q}$	Vertex random effect variance for the qth task.
$\boldsymbol{\Sigma}$	$NVQ \times NVQ$ covariance matrix of $\boldsymbol{d}$.
t	Index for time.
t_0	Time-delay parameter of canonical HRF.
T	Total number of timepoints.
τ^2_{nv}	Innovation variance of the AR errors for the nth subject at the vth voxel.
θ_q	Spatial dependence parameter for the qth task.
u_{vq}	Vertex random effect at the vth vertex for the qth task.
$\boldsymbol{u}^q$	$[u_{1q}, \ldots, u_{Vq}]'$
$\boldsymbol{u}_v$	$[u_{v1}, \ldots, u_{vQ}]'$
$\boldsymbol{u}$	$[\boldsymbol{u}_1', \ldots, \boldsymbol{u}_V']'$
$\mathbf{U}$	$\text{diag}(\sigma^2_{u_1}, \ldots, \sigma^2_{u_Q})$
v	Index for vertex.
V	Total number of vertices. In Section 3.3, this is the total number of vertices in a region, where the index on region has been dropped for succinctness.
$\mathcal{V}_r$	The set of vertices, $\{v_1, \ldots, v_{V_r}\}$, in the rth region.
$\mathbf{W}$	$\text{diag}(w_1, \ldots, w_Q)$, where w_q is the sum of all elements in $\boldsymbol{\Omega}_q$.
x_{ntq}	The qth covariate of the nth subject at the tth timepoint.
$\mathbf{x}_{nt}$	$[x_{nt1}, \ldots, x_{ntQ}]'$
$\mathbf{X}_n$	$[\mathbf{x}_{n1}, \ldots, \mathbf{x}_{nT}]'$
$\mathbf{X}^*_n$	The first-level design matrix with nuisance covariates: $[\mathbf{X}_n, \mathbf{Z}_n]$.
ξ^2_{nv}	Unconditional variance of the AR errors
y_{nvt}	The BOLD signal of the nth subject measured at the vth voxel and tth timepoint.
$\mathbf{Y}_{nv}$	$[y_{nv1}, \ldots, y_{nvT}]'$
$\mathbf{Y}_n$	$[y_{n11}, \ldots, y_{n1T}, y_{n21}, \ldots, y_{n2T}, \ldots, y_{nVT}]'$
$\mathbf{Y}$	$[\mathbf{Y}_1,', \ldots, \mathbf{Y}_N']'$
z_{nvtm}	The mth nuisance covariate of the nth subject, at the tth timepoint.

Continued on next page

Table C.1 – continued from previous page

Notation	Description
$\mathbf{z}_{nt}$	$[z_{nt1}, \ldots, z_{ntM}]'$
$\mathbf{Z}_n$	$\{_r\ \mathbf{z}'_{nt}\}_{t=1}^{T}$
ζ_0	Dispersion parameter of the canonical HRF.

C.2 Accounting for uncertainty in the timing and duration of the HRF

Activation models can include a covariate equal to the partial derivative of the HRF with respect to the time-delay parameter and a covariate for the partial derivative with respect to a dispersion parameter, which allows the temporal delay after stimulus and the length of the response to a stimulus to vary spatially (Chapter 12, Frackowiak et al. 2004). Including these derivatives results in a first-order Taylor series approximation to the HRF for a particular vertex about the canonical time-delay and dispersion parameters. Let η_{nv} and ζ_{nv} denote the true time-delay and dispersion parameters, respectively, for the nth subject at the vth vertex. Let η_0 and ζ_0 denote the time-delay and dispersion parameters in the canonical HRF. Then let $x_{ntq}(\eta, \zeta)$ be a function of η and ζ resulting from convolving the qth task with the canonical HRF. Define $\Delta\eta_{nv} = \eta_{nv} - \eta_0$ and $\Delta\zeta_{nv} = \zeta_{nv} - \zeta_0$. Then consider the first-order Taylor series approximation of $x_{ntq}(\eta_{nv}, \zeta_{nv})$ about the canonical parameters:

$$\begin{aligned} x_{ntq}(\eta_{nv}, \zeta_{nv}) &= x_{ntq}(\eta_0 + \Delta\eta_{nv}, \zeta_0 + \Delta\zeta_{nv}) \\ &= x_{ntq}(\eta_0, \zeta_0) + \Delta\eta_{nv}\frac{\partial x_{ntq}}{\partial \eta}(\eta_0, \zeta_0) + \Delta\zeta_{nv}\frac{\partial x_{ntq}}{\partial \zeta}(\eta_0, \zeta_0) + o(\|\Delta\eta_{nv}, \Delta\zeta_{nv}\|), \end{aligned}$$

where $\frac{\partial x_{ntq}}{\partial \eta}(\eta_0, \zeta_0)$ denotes the partial derivative with respect to η evaluated at η_0 and ζ_0.

To keep the exposition simple and for concreteness, let us assume for the moment that there is one task and no nuisance covariates. We would like to estimate the following model:

$$y_{nvt} = d_{nv1} x_{nt1}(\eta_{nv}, \zeta_{nv}) + a^*_{nvt}$$

where a^*_{nvt} are the errors when the exact HRF is used. Define

$$a_{nvt} = a^*_{nvt} + \left\{ x_{nt1}(\eta_{nv}, \zeta_{nv}) - x_{nt1}(\eta_0, \zeta_0) - \Delta\eta_{nv} \frac{\partial x_{nt1}}{\partial \eta}(\eta_0, \zeta_0) - \Delta\zeta_{nv} \frac{\partial x_{nt1}}{\partial \zeta}(\eta_0, \zeta_0) \right\}.$$

Then we have

$$y_{nvt} = d_{nv1} x_{nt1}(\eta_0, \zeta_0) + d_{nv1}\Delta\eta_{nv} \frac{\partial x_{nt1}}{\partial \eta}(\eta_0, \zeta_0) + d_{nv1}\Delta\zeta_{nv} \frac{\partial x_{nt1}}{\partial \zeta}(\eta_0, \zeta_0) + a_{nvt}.$$

The linear model in (3.1) specifies a different parameterization:

$$y_{nvt} = d_{nv1} x_{nt1}(\eta_0, \zeta_0) + \gamma_{nv1} \frac{\partial x_{ntq}}{\partial \eta}(\eta_0, \zeta_0) + \gamma_{nv2} \frac{\partial x_{ntq}}{\partial \zeta}(\eta_0, \zeta_0) + a_{nvt}$$

where $\gamma_{nv1} = d_{nv1}\Delta\eta_{nv}$ and $\gamma_{nv2} = d_{nv1}\Delta\zeta_{nv}$,

One could construct F-tests to assess the overall effect of a task taking into account the time-delay and dispersal derivatives, although such an F-test would not provide information on the sign of the overall activation. In this paper, we treat the time-delay and dispersal derivatives as nuisance covariates. Their inclusion reduces the bias in estimates of d_{nvq}.

C.3 Biasedness of the OLS estimator of the error variance

Consider the OLS estimator of the error variance:

$$\hat{\xi}^2_{OLS,nv} = \frac{1}{T - (Q + M)} \sum_{t=1}^{T} (y_{nvt} - \hat{y}_{nvt})^2.$$

Define $\widehat{\mathbf{Y}}_{nv} = \mathbf{H}_n \mathbf{Y}_{nv}$. Let $\hat{y}_{nvt}$ be the corresponding element of $\widehat{\mathbf{Y}}_{nv}$. Note that

$$\frac{1}{T-(Q+M)} \mathrm{E} \sum_{t=1}^{T} (y_{nvt} - \hat{y}_{nvt})^2 = \frac{1}{T-(Q+M)} \mathrm{E} \ \mathrm{tr} \sum_{t=1}^{T} (y_{nvt} - \hat{y}_{nvt})^2$$
$$= \frac{1}{T-(Q+M)} \mathrm{tr} \, \mathrm{E} \, (\mathbf{Y}_{nv} - \mathbf{H}_n \mathbf{Y}_{nv})(\mathbf{Y}_{nv} - \mathbf{H}_n \mathbf{Y}_{nv})'.$$

We also have $\mathrm{E} \, \mathbf{H}_n \mathbf{Y}_{nv} = \mathbf{X}_n^* (\mathbf{X}_n^{*\prime} \mathbf{X}_n^*)^{-1} \mathbf{X}_n^{*\prime} [\boldsymbol{\beta}', \boldsymbol{\gamma}_{nv}']' = \mathbf{E}\mathbf{Y}_{nv}$. Then write

$$\mathrm{E} \, (\mathbf{Y}_{nv} - \mathbf{H}_n \mathbf{Y}_{nv})(\mathbf{Y}_{nv}' - \mathbf{Y}_{nv}' \mathbf{H}_n) =$$
$$\mathrm{Cov} \, \mathbf{Y}_{nv} - \mathrm{Cov} \, (\mathbf{Y}_{nv}, \mathbf{H}_n \mathbf{Y}_{nv}) - \mathrm{Cov} \, (\mathbf{H}_n \mathbf{Y}_{nv}, \mathbf{Y}_{nv}) + \mathrm{Cov} \, (\mathbf{H}_n \mathbf{Y}_{nv}). \tag{C.1}$$

Now, $\mathrm{Cov} \, \mathbf{Y}_{nv} = \mathbf{X}_n (\mathbf{U} + \mathbf{S} + \mathbf{B}) \mathbf{X}_n' + \boldsymbol{\Psi}_{nv}$. Note $\mathbf{H}_n \mathbf{X}_n = \mathbf{X}_n$ because $\mathbf{X}_n$ is in the column space of $\mathbf{H}_n$. The term $\mathbf{X}_n (\mathbf{U} + \mathbf{S} + \mathbf{B}) \mathbf{X}_n'$ appears in each of the four covariance terms in (C.1), and thus drops out. Then we are left with

$$= \boldsymbol{\Psi}_{nv} - \boldsymbol{\Psi}_{nv} \mathbf{H}_n - \mathbf{H}_n \boldsymbol{\Psi}_{nv} + \mathbf{H}_n \boldsymbol{\Psi}_{nv} \mathbf{H}_n$$
$$= (\mathbf{I}_T - \mathbf{H}_n) \boldsymbol{\Psi}_{nv} (\mathbf{I}_T - \mathbf{H}_n). \tag{C.2}$$

Note that $(\mathbf{I}_n - \mathbf{H}_n)(\mathbf{I}_n - \mathbf{H}_n) = \mathbf{I}_n - \mathbf{H}_n$. Then we have

$$\mathrm{tr} \ \{(\mathbf{I}_T - \mathbf{H}_n) \boldsymbol{\Psi}_{nv} (\mathbf{I}_T - \mathbf{H}_n)\} = \mathrm{tr} \ \{\boldsymbol{\Psi}_{nv} (\mathbf{I}_T - \mathbf{H}_n)\}$$
$$= T\sigma_{nv}^2 - \mathrm{tr} \, (\boldsymbol{\Psi}_{nv} \mathbf{H}_n).$$

from which (3.13) follows.

C.4 Deriving the expected value of the MSB

We can derive estimators for a vector response $\boldsymbol{d}_{nv} \in \mathbb{R}^Q$, corresponding to multiple tasks. Below, we provide details of this calculation, although it is somewhat tedious.

We will utilize the following quantities:

$$\text{Cov}\left(\boldsymbol{d}_{nv}, \bar{\boldsymbol{d}}_{n\cdot}\right) = \frac{1}{V}\sum_{v'=1}^{V}\boldsymbol{\Gamma}_{v,v'}\mathbf{U} + \mathbf{S} + \frac{1}{V}\sum_{v'=1}^{V}\boldsymbol{\Omega}_{v,v'}\mathbf{B} + \frac{1}{V}\xi_{nv}^2\mathbf{K}_n'\boldsymbol{\Psi}_{nv}\mathbf{K}_n; \tag{C.3}$$

$$\text{Cov}\left(\boldsymbol{d}_{nv}, \bar{\boldsymbol{d}}_{\cdot v}\right) = \mathbf{U} + \frac{1}{N}\mathbf{S} + \frac{1}{N}\mathbf{B} + \frac{1}{N}\xi_{nv}^2\mathbf{K}_n'\boldsymbol{\Psi}_{nv}\mathbf{K}_n;$$

$$\text{Cov}\left(\boldsymbol{d}_{nv}, \bar{\boldsymbol{d}}_{\cdot\cdot}\right) = \frac{1}{V}\sum_{v'=1}^{V}\boldsymbol{\Gamma}_{v,v'}\mathbf{U} + \frac{1}{N}\mathbf{S} + \frac{1}{NV}\sum_{v'=1}^{V}\boldsymbol{\Omega}_{v,v'}\mathbf{B} + \frac{1}{NV}\xi_{nv}^2\mathbf{K}_n'\boldsymbol{\Psi}_{nv}\mathbf{K}_n;$$

$$\text{Cov}\left(\bar{\boldsymbol{d}}_{n\cdot}, \bar{\boldsymbol{d}}_{\cdot v}\right) = \frac{1}{V}\sum_{v'=1}^{V}\boldsymbol{\Gamma}_{v,v'}\mathbf{U} + \frac{1}{N}\mathbf{S} + \frac{1}{NV}\sum_{v'=1}^{V}\boldsymbol{\Omega}_{v,v'}\mathbf{B} + \frac{1}{NV}\xi_{nv}^2\mathbf{K}_n'\boldsymbol{\Psi}_{nv}\mathbf{K}_n;$$

$$\text{Cov}\left(\bar{\boldsymbol{d}}_{n\cdot}\right) = \mathbf{S} + \frac{1}{V^2}\mathbf{G}\mathbf{U} + \frac{1}{V^2}\mathbf{W}\mathbf{B} + \frac{1}{V^2}\sum_{v=1}^{V}\xi_{nv}^2\mathbf{K}_n'\boldsymbol{\Psi}_{nv}\mathbf{K}_n;$$

$$\text{Cov}\left(\bar{\boldsymbol{d}}_{n\cdot}, \bar{\boldsymbol{d}}_{\cdot\cdot}\right) = \frac{1}{N}\mathbf{S} + \frac{1}{V^2}\mathbf{G}\mathbf{U} + \frac{1}{NV^2}\mathbf{W}\mathbf{B} + \frac{1}{NV^2}\sum_{v=1}^{V}\xi_{nv}^2\mathbf{K}_n'\boldsymbol{\Psi}_{nv}\mathbf{K}_n;$$

$$\text{Cov}\left(\bar{\boldsymbol{d}}_{\cdot v}\right) = \mathbf{U} + \frac{1}{N}\mathbf{S} + \frac{1}{N}\mathbf{B} + \frac{1}{N^2}\sum_{n=1}^{N}\xi_{nv}^2\mathbf{K}_n'\boldsymbol{\Psi}_{nv}\mathbf{K}_n;$$

$$\text{Cov}\left(\bar{\boldsymbol{d}}_{\cdot v}, \bar{\boldsymbol{d}}_{\cdot\cdot}\right) = \frac{1}{V}\mathbf{U} + \frac{1}{V}\sum_{v'\neq v}\boldsymbol{\Gamma}_{v,v'}\mathbf{U} + \frac{1}{N}\mathbf{S} + \frac{1}{NV}\sum_{v'=1}^{V}\boldsymbol{\Omega}_{v,v'}\mathbf{B} + \frac{1}{N^2V}\sum_{n=1}^{N}\xi_{nv}^2\mathbf{K}_n'\boldsymbol{\Psi}_{nv}\mathbf{K}_n; \text{ and}$$

$$\text{Cov}\left(\bar{\boldsymbol{d}}_{\cdot\cdot}\right) = \frac{1}{N}\mathbf{S} + \frac{1}{V^2}\mathbf{G}\mathbf{U} + \frac{1}{NV^2}\mathbf{W}\mathbf{B} + \frac{1}{N^2V^2}\sum_{n=1}^{N}\sum_{v=1}^{V}\xi_{nv}^2\mathbf{K}_n'\boldsymbol{\Psi}_{nv}\mathbf{K}_n.$$

Note $\mathrm{E}\,\bar{\boldsymbol{d}}_{n\cdot} = \mathrm{E}\,\bar{\boldsymbol{d}}_{\cdot v} = \mathrm{E}\,\bar{\boldsymbol{d}}_{\cdot\cdot}$. Then,

$$
\begin{aligned}
&\mathrm{E}\,(\boldsymbol{d}_{nv} - \bar{\boldsymbol{d}}_{n\cdot} - \bar{\boldsymbol{d}}_{\cdot v} + \bar{\boldsymbol{d}}_{\cdot\cdot})(\boldsymbol{d}_{nv} - \bar{\boldsymbol{d}}_{n\cdot} - \bar{\boldsymbol{d}}_{\cdot v} + \bar{\boldsymbol{d}}_{\cdot\cdot})' \\
&= \mathrm{Cov}\,\boldsymbol{d}_{nv} + \mathrm{Cov}\,\bar{\boldsymbol{d}}_{n\cdot} + \mathrm{Cov}\,\bar{\boldsymbol{d}}_{\cdot v} + \mathrm{Cov}\,\bar{\boldsymbol{d}}_{\cdot\cdot} - 2\mathrm{Cov}\,(\boldsymbol{d}_{nv}, \bar{\boldsymbol{d}}_{n\cdot}) - 2\mathrm{Cov}\,(\boldsymbol{d}_{nv}, \bar{\boldsymbol{d}}_{\cdot v}) \\
&\quad + 2\mathrm{Cov}\,(\boldsymbol{d}_{nv}, \bar{\boldsymbol{d}}_{\cdot\cdot}) + 2\mathrm{Cov}\,(\bar{\boldsymbol{d}}_{n\cdot}, \bar{\boldsymbol{d}}_{v\cdot}) - 2\mathrm{Cov}\,(\bar{\boldsymbol{d}}_{n\cdot}, \bar{\boldsymbol{d}}_{\cdot\cdot}) - 2\mathrm{Cov}\,(\bar{\boldsymbol{d}}_{v\cdot}, \bar{\boldsymbol{d}}_{\cdot\cdot}) \\
&= \mathbf{U} + \mathbf{S} + \mathbf{B} + \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n \\
&+ \frac{1}{V^2}\mathbf{G}\mathbf{U} + \mathbf{S} + \frac{1}{V^2}\mathbf{W}\mathbf{B} + \frac{1}{V^2}\sum_{v=1}^{V} \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n; \\
&+ \mathbf{U} + \frac{1}{N}\mathbf{S} + \frac{1}{N}\mathbf{B} + \frac{1}{N^2}\sum_{n=1}^{N} \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n \\
&+ \frac{1}{V^2}\mathbf{G}\mathbf{U} + \frac{1}{N}\mathbf{S} + \frac{1}{NV^2}\mathbf{W}\mathbf{B} + \frac{1}{N^2V^2}\sum_{n=1}^{N}\sum_{v=1}^{V} \mathbf{K}_n \boldsymbol{\Psi}_{nv} \mathbf{K}_n' \\
&- \frac{2}{V}\sum_{v'=1}^{V} \boldsymbol{\Gamma}_{v,v'}\mathbf{U} - 2\mathbf{S} - \frac{2}{V}\sum_{v'=1}^{V} \boldsymbol{\Omega}_{v,v'}\mathbf{B} - \frac{2}{V}\xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n \\
&- 2\mathbf{U} - \frac{2}{N}\mathbf{S} - \frac{2}{N}\mathbf{B} - \frac{2}{N}\xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n \\
&+ \frac{4}{V}\sum_{v'=1}^{V} \boldsymbol{\Gamma}_{v,v'}\mathbf{U} + \frac{4}{N}\mathbf{S} + \frac{4}{NV}\sum_{v'=1}^{V} \boldsymbol{\Omega}_{v,v'}\mathbf{B} + \frac{4}{NV}\xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n \\
&- \frac{2}{V^2}\mathbf{G}\mathbf{U} - \frac{2}{N}\mathbf{S} - \frac{2}{NV^2}\mathbf{W}\mathbf{B} - \frac{2}{NV^2}\sum_{v=1}^{V} \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n \\
&- \frac{2}{V}\sum_{v'\neq v} \boldsymbol{\Gamma}_{v,v'}\mathbf{U} - \frac{2}{N}\mathbf{S} - \frac{2}{NV}\sum_{v'=1}^{V} \boldsymbol{\Omega}_{v,v'}\mathbf{B} - \frac{2}{N^2V}\sum_{n=1}^{N} \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n.
\end{aligned}
$$

The terms involving **U** and **S** drop out,

$$
\begin{aligned}
&= \mathbf{B} + \xi_{nv}^2 \mathbf{K}_n' \mathbf{\Psi}_{nv} \mathbf{K}_n \\
&+ \frac{1}{V^2}\mathbf{WB} + \frac{1}{V^2}\sum_{v=1}^{V} \xi_{nv}^2 \mathbf{K}_n' \mathbf{\Psi}_{nv} \mathbf{K}_n; \\
&+ \frac{1}{N}\mathbf{B} + \frac{1}{N^2}\sum_{n=1}^{N} \xi_{nv}^2 \mathbf{K}_n' \mathbf{\Psi}_{nv} \mathbf{K}_n \\
&+ \frac{1}{NV^2}\mathbf{WB} + \frac{1}{N^2V^2}\sum_{n=1}^{N}\sum_{v=1}^{V} \mathbf{K}_n \mathbf{\Psi}_{nv} \mathbf{K}_n' \\
&- \frac{2}{V}\sum_{v'=1}^{V} \mathbf{\Omega}_{v,v'}\mathbf{B} - \frac{2}{V}\xi_{nv}^2 \mathbf{K}_n' \mathbf{\Psi}_{nv} \mathbf{K}_n \\
&- \frac{2}{N}\mathbf{B} - \frac{2}{N}\xi_{nv}^2 \mathbf{K}_n' \mathbf{\Psi}_{nv} \mathbf{K}_n \\
&+ \frac{4}{NV}\sum_{v'=1}^{V} \mathbf{\Omega}_{v,v'}\mathbf{B} + \frac{4}{NV}\xi_{nv}^2 \mathbf{K}_n' \mathbf{\Psi}_{nv} \mathbf{K}_n \\
&- \frac{2}{NV^2}\mathbf{WB} - \frac{2}{NV^2}\sum_{v=1}^{V} \xi_{nv}^2 \mathbf{K}_n' \mathbf{\Psi}_{nv} \mathbf{K}_n \\
&- \frac{2}{NV}\sum_{v'=1}^{V} \mathbf{\Omega}_{v,v'}\mathbf{B} - \frac{2}{N^2V}\sum_{n=1}^{N} \xi_{nv}^2 \mathbf{K}_n' \mathbf{\Psi}_{nv} \mathbf{K}_n.
\end{aligned}
$$

Putting this together,

$$
\mathrm{E}\, MSB = \mathbf{B} - \frac{2}{V(V-1)}\sum_{v=1}^{V-1}\sum_{v'=v+1}^{V} \mathbf{\Omega}_{v,v'}\mathbf{B} + \frac{1}{NV}\sum_{n=1}^{N}\sum_{v=1}^{V} \xi_{nv}^2 \mathbf{K}_n' \mathbf{\Psi}_{nv} \mathbf{K}_n.
$$

which is equivalent to (3.16).

C.5 Satterthwaite-like approximation to the degrees of freedom

Here, we describe an approach whose performance will be investigated in future research. Consider the case where we are testing the significance of a main effect, $H_0 : \beta_q = 0$. As the number of subjects grows, this test statistic approaches a stan-

dard normal distribution. Since the number of subjects in fMRI studies can be small, we would like a more accurate approximation.

The Satterthwaite approach provides an approximate distribution for a random variable that is a linear combination of independent chi-squared variables, where the distribution is approximated by a chi-squared variable with the same mean and variance (Satterthwaite, 1946). In its usual application, the expected value of this linear combination is equal to the variance due to a fixed or random factor under the null. We can not directly use the Satterthwaite approach to approximate the distribution of our estimate of (3.24) because it is a non-linear combination of the variance components due to the inverse. As an alternative approach, we derive the linear combination of variance components that determine the variance of $\bar{\boldsymbol{d}}_{\cdot\cdot}$, which we will use as a surrogate to the approximate degrees of freedom for our actual estimator. Consider the mean square of the overall mean:

$$\begin{aligned}
\mathrm{E} \sum_{n=1}^{N} \sum_{v=1}^{V} \bar{\boldsymbol{d}}_{\cdot\cdot} \bar{\boldsymbol{d}}_{\cdot\cdot}' &= NV \mathrm{Cov}\, \bar{\boldsymbol{d}}_{\cdot\cdot} + NV \boldsymbol{\beta}\boldsymbol{\beta}' \\
&= V\mathbf{S} + \frac{N}{V}\mathbf{GU} + \frac{1}{V}\mathbf{WB} + \frac{1}{NV} \sum_{n=1}^{N} \sum_{v=1}^{V} \xi_{nv}^2 \mathbf{K}_n' \boldsymbol{\Psi}_{nv} \mathbf{K}_n + \boldsymbol{\beta}\boldsymbol{\beta}' \\
&= \mathrm{E}\, MSS + \frac{N}{V}\mathbf{GU} + NV\boldsymbol{\beta}\boldsymbol{\beta}'. \qquad \text{(C.4)}
\end{aligned}$$

Consider the case for $Q = 1$. Then from (3.18), we have

$$\sigma_u^2 = \frac{V(V-1)}{N(V^2 - g)} \left(\mathrm{E}\, MSU - \mathrm{E}\, MSB\right)$$

and it follows that

$$\frac{N}{V} g \sigma_u^2 = \frac{V-1}{V^2/g - 1} \left(\mathrm{E}\, MSU - \mathrm{E}\, MSB\right),$$

so we have

$$\mathrm{E} \sum_{n=1}^{N} \sum_{v=1}^{V} \bar{d}_{\cdot\cdot}^2 = \mathrm{E}\, MSS + \frac{V-1}{V^2/g - 1} \left(\mathrm{E}\, MSU - \mathrm{E}\, MSB\right) + NV\beta^2.$$

If we had the usual two-factor crossed design for a univariate response, then to construct an approximate F-statistic to test whether the fixed effect is significantly different from zero, one would take the ratio of the mean squares from the intercept and the above equation with expectations replaced by their sample approximations. Here, the degrees of freedom for the numerator is equal to one. Let

$$\alpha = \left(\frac{V-1}{V^2/\hat{g} - 1} \right).$$

An approximate degrees of freedom for the denominator is then

$$\widehat{df} = \frac{(MSS + \alpha MSU - \alpha MSB)^2}{MSS^2/(N-1) + \alpha^2 MSU^2/(V-1) + \alpha^2 MSB^2/\{(N-1)(V-1)\}}. \tag{C.5}$$

Rather than construct the F-test based on $\sum\sum \bar{d}_{..}^2$, we will use (C.5) to approximate the degrees of freedom in (3.28). It should be noted that this is a departure from the standard approach. It is usually the case that $\hat{\beta} = \bar{d}_{..}$, in which case the square of the Wald t-statistic is equal to the F-statistic with the denominator of the F-statistic equal to the variance of the estimator divided by its degrees of freedom; but here, our estimator is the GLS estimator, and the variance of the GLS estimator is not a linear combination of the mean squares.

For $q > 1$, we can calculate the above quantity for each q. For calculating a contrast between two tasks, we average the two estimates.

C.6 Covariates included in the analysis of ToM HCP data

	Covariate	Description
1	xMental	HRF convolved with mentalizing task
2	xRandom	HRF convolved with random task
3	zIntercept	Intercept
4	z_dxMentaldt	Derivative of 'xMental' with respect to time-delay parameter
5	z_dxMentaldd	Derivative of 'xMental' with respect to dispersion parameter
6	z_dxRandomdt	Derivative of 'xRandom' with respect to time-delay parameter
7	z_dxRandomdd	Derivative of 'xRandom' with respect to dispersion parameter
8	zSession	Indicator variable for session
9	zBasis1_sess1	First basis for the piece-wise linear spline indicating the time in seconds corresponding to each fMRI volume from the first session (from 0 to 197 seconds) and equal to zero during the second session
10	zBasis2_sess1	Second basis corresponding to a knot at 49 seconds, equal to zero after 197 seconds
11	zBasis3_sess1	Third basis with knot at 99 seconds, equal to zero after 197 seconds
12	zBasis4_sess1	Fourth basis with knot at 148 seconds, equal to zero after 197 seconds
13	zBasis1_sess2	First basis for session 2 equal to zero during the first session and counting from zero starting from the 275th time point
14	zBasis2_sess2	Second basis for session two
15	zBasis3_sess2	Third basis for session two
16	zBasis4_sess2	Fourth basis for session two
17	zTransX_sess1	Subject-specific motion parameter from affine registration of first session: shift in x-coordinate
18	zTransY_sess1	Subject-specific motion parameter from affine registration of first session: shift in y-coordinate
19	zTransZ_sess1	Subject-specific motion parameter from affine registration of first session: shift in z-coordinate
20	zRotX_sess1	Subject-specific motion parameter from affine registration of first session: rotation in x-coordinate
21	zRotY_sess1	: rotation in y-coordinate
22	zRotZ_sess1	: rotation in z-coordinate
23	zTransX_sess2	Subject-specific motion parameter from affine registration of second session: shift in x-coordinate
24	zTransY_sess2	: shift in y-coordinate
25	zTransZ_sess2	: shift in z-coordinate
26	zRotX_sess2	: rotation in x-coordinate
27	zRotY_sess2	: rotation in y-coordinate
28	zRotZ_sess2	: rotation in z-coordinate

Table C.2: Covariates included in the HCP ToM analysis. Note that 'xMental' and 'xRandom' are the covariates of interest (composing $\mathbf{X}$) and the others are nuisance covariates (composing $\mathbf{Z}$).

BIBLIOGRAPHY

Allassonniere, S. and Younes, L. (2012). A stochastic algorithm for probabilistic independent component analysis. *The Annals of Applied Statistics*, 6(1):125–160.

Amari, S. I., Cichocki, A., and Yang, H. H. (1996). A new learning algorithm for blind signal separation. *Advances in neural information processing systems*, pages 757–763.

Amato, U., Antoniadis, A., Samarov, A., and Tsybakov, A. (2010). Noisy independent factor analysis model for density estimation and classification. *Electronic Journal of Statistics*, 4:707–736.

Attias, H. (1999). Independent factor analysis. *Neural computation*, 11(4):803–851.

Bach, F. R. and Jordan, M. I. (2003). Kernel independent component analysis. *The Journal of Machine Learning Research*, 3:1–48.

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., et al. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, 80:169–189.

Bartlett, M. S., Movellan, J. R., and Sejnowski, T. J. (2002). Face recognition by independent component analysis. *Neural Networks, IEEE Transactions on*, 13(6):1450–1464.

Beckmann, C. F. (2012). Modelling with independent components. *NeuroImage*, 62(2):891–901.

Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2003). General multilevel linear modeling for group analysis in fMRI. *Neuroimage*, 20(2):1052–1063.

Beckmann, C. F. and Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging*, 23(2):137–152.

Beckmann, C. F. and Smith, S. M. (2005). Tensorial extensions of independent component analysis for multisubject fMRI analysis. *Neuroimage*, 25(1):294–311.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, pages 289–300.

Bernaards, C. A. and Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement*, 65(5):676–696.

Bernal-Rusiel, J. L., Reuter, M., Greve, D. N., Fischl, B., and Sabuncu, M. R. (2013). Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data. *NeuroImage*, 81:358–370.

Biswal, B. B., Mennes, M., Zuo, X. N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., et al. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739.

Blanchard, G., Kawanabe, M., Sugiyama, M., Spokoiny, V., and Müller, K.-R. (2006). In search of non-Gaussian components of a high-dimensional distribution. *The Journal of Machine Learning Research*, 7:247–282.

Bowman, F. D. (2005). Spatio-temporal modeling of localized brain activity. *Biostatistics*, 6(4):558–575.

Bowman, F. D. (2007). Spatiotemporal models for region of interest analyses of functional neuroimaging data. *Journal of the American Statistical Association*, 102(478):442–453.

Calhoun, V. D., Adali, T., Pearlson, G. D., and Pekar, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human brain mapping*, 14(3):140–151.

Calhoun, V. D., Liu, J., and Adali, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage*, 45(1):S163–S172.

Cardoso, J. F. (1997). Infomax and maximum likelihood for blind source separation. *Signal Processing Letters IEEE*, 4(4):112–114.

Cardoso, J. F. (1998). Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025.

Cardoso, J. F. and Souloumiac, A. (1993). Blind beamforming for non-Gaussian signals. In *Radar and Signal Processing, IEEE Proceedings F*, volume 140, pages 362–370.

Celone, K. A., Calhoun, V. D., Dickerson, B. C., Atri, A., Chua, E. F., Miller, S. L., DePeau, K., Rentz, D. M., Selkoe, D. J., Blacker, D., et al. (2006). Alterations in memory networks in mild cognitive impairment and Alzheimer's disease: an independent component analysis. *The Journal of neuroscience*, 26(40):10222–10231.

Chen, A. and Bickel, P. J. (2006). Efficient independent component analysis. *The Annals of Statistics*, 34(6):2825–2855.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.

Correa, N., Adali, T., and Calhoun, V. D. (2007). Performance of blind source separation algorithms for fMRI analysis using a group ICA method. *Magnetic resonance imaging*, 25(5):684–694.

Damoiseaux, J. S., Rombouts, S., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., and Beckmann, C. F. (2006). Consistent resting-state networks across healthy subjects. *Proceedings of the national academy of sciences*, 103(37):13848–13853.

Delorme, A., Sejnowski, T., and Makeig, S. (2007). Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *Neuroimage*, 34(4):1443–1449.

Derado, G., Bowman, F. D., and Kilts, C. D. (2010). Modeling the spatial and temporal dependence in fMRI data. *Biometrics*, 66(3):949–957.

Derado, G., Bowman, F. D., Zhang, L., et al. (2013). Predicting brain activity using a Bayesian spatial model. *Statistical methods in medical research*, 22(4):382–397.

Eloyan, A. and Ghosh, S. K. (2013). A semiparametric approach to source separation using independent component analysis. *Computational Statistics and Data Analysis*, 58:383 – 396.

Eloyan, A., Muschelli, J., Nebel, M. B., Liu, H., Han, F., Zhao, T., Barber, A., Joel, S., Pekar, J. J., Mostofsky, S., and Caffo, B. S. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in Systems Neuroscience*, 6:1–9.

Fischl, B., Sereno, M. I., Tootell, R. B., Dale, A. M., et al. (1999). High-resolution

intersubject averaging and a coordinate system for the cortical surface. *Human brain mapping*, 8(4):272–284.

Frackowiak, R. S., Friston, K. J., Frith, C. D., Dolan, R. J., Price, C. J., Zeki, S., Ashburner, J. T., and Penny, W. D. (2004). *Human Brain Function*. Academic Press, 2nd edition.

Friston, K. J., Stephan, K. E., Lund, T. E., Morcom, A., and Kiebel, S. (2005). Mixed-effects and fMRI studies. *Neuroimage*, 24(1):244–252.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., et al. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage*, 80:105–124.

Gordon, E. M., Laumann, T. O., Adeyemo, B., Huckins, J. F., Kelley, W. M., and Petersen, S. E. (2014). Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral Cortex*, page bhu239.

Green, C. G., Nandy, R. R., and Cordes, D. (2002). PCA-preprocessing of fMRI data adversely affects the results of ICA. In *Proceedings of international society of magnetic resonance in medicine*, page 10.

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., Zsoldos, E., Ebmeier, K. P., Filippini, N., Mackay, C. E., et al. (2014). ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *NeuroImage*, 95:232–247.

Guo, Y. (2011). A general probabilistic model for group independent component analysis and its estimation methods. *Biometrics*, 67:1532–1542.

Guo, Y. and Tang, L. (2013). A hierarchical model for probabilistic independent component analysis of multi-subject fMRI studies. *Biometrics*, 69(4):970–981.

Harrison, L. M. and Green, G. G. (2010). A Bayesian spatiotemporal model for very large data sets. *NeuroImage*, 50(3):1126–1141.

Hastie, T. (2013). *GAM: Generalized Additive Models*. R package version 1.08.

Hastie, T. and Tibshirani, R. (2003). Independent components analysis through product density estimation. *Advances in Neural Information Processing Systems*, 15:649–656.

Hastie, T. and Tibshirani, R. (2010). *ProDenICA: Product Density Estimation for ICA using tilted Gaussian density estimates*. R package version 1.0.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer.

Himberg, J., Hyvärinen, A., and Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. *Neuroimage*, 22(3):1214–1222.

Holmes, A. and Friston, K. (1998). Generalisability, random effects, and population inference. In *Proceedings of the Fourth International Conference on Functional Mapping of the Human Brain, June 7-12, Montreal, Canada.*, number S754.

Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, pages 435–475.

Hyun, J. W., Li, Y., Gilmore, J. H., Lu, Z., Styner, M., and Zhu, H. (2014). SGPP: spatial gaussian predictive process models for neuroimaging data. *NeuroImage*, 89:70–80.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634.

Hyvärinen, A., Hoyer, P., and Oja, E. (1999). Image denoising by sparse code shrinkage. In *Intelligent Signal Processing*. Citeseer.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent component analysis*. Wiley-Interscience.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430.

Ilmonen, P., Nordhausen, K., Oja, H., and Ollila, E. (2010). A new performance index for ICA: properties, computation and asymptotic analysis. *Latent Variable Analysis and Signal Separation*, pages 229–236.

Iriarte, J., Urrestarazu, E., Valencia, M., Alegre, M., Malanda, A., Viteri, C., and Artieda, J. (2003). Independent component analysis as a tool to eliminate artifacts in EEG: a quantitative study. *Journal of clinical neurophysiology*, 20(4):249–257.

Jafri, M. J., Pearlson, G. D., Stevens, M., and Calhoun, V. D. (2008). A method for functional network connectivity among spatially independent resting-state components in schizophrenia. *Neuroimage*, 39(4):1666–1681.

Kagan, A. M., Rao, C. R., and Linnik, Y. V. (1973). *Characterization problems in mathematical statistics*. Wiley.

Kang, H., Ombao, H., Linkletter, C., Long, N., and Badre, D. (2012). Spatio-spectral mixed-effects model for functional magnetic resonance imaging data. *Journal of the American Statistical Association*, 107(498):568–577.

Kawanabe, M., Sugiyama, M., Blanchard, G., and Müller, K. (2007). A new algorithm of non-Gaussian component analysis with radial kernel functions. *Annals of the Institute of Statistical Mathematics*, 59(1):57–75.

Kong, W., Vanderburg, C. R., Gunshin, H., Rogers, J. T., and Huang, X. (2008). A review of independent component analysis application to microarray gene expression data. *Biotechniques*, 45(5):501.

Kuhn, H. W. (1955). The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83 – 97.

Lee, T. W., Girolami, M., and Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural computation*, 11(2):417–441.

Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Statistical Science*, 23(4):439–464.

Marchini, J. L., Heaton, C., and Ripley, B. D. (2010). *FastICA: FastICA Algorithms to perform ICA and Projection Pursuit*. R package version 1.1-13.

Matteson, D. S. and Tsay, R. S. (2013). Independent component analysis via distance covariance. *ArXiv e-prints*.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized linear mixed models*. Wiley, second edition.

Mennes, M., Biswal, B., Castellanos, F. X., and Milham, M. P. (2012). Making data sharing work: The FCP/INDI experience. *NeuroImage*.

Mikl, M., Mareček, R., Hluštík, P., Pavlicová, M., Drastich, A., Chlebus, P., Brázdil, M., and Krupa, P. (2008). Effects of spatial smoothing on fMRI group inferences. *Magnetic resonance imaging*, 26(4):490–503.

Milham, M. P., Fair, D., Mennes, M., and Mostofsky, S. H. (2012). The ADHD-200

consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, 6:62.

Mumford, J. A. and Nichols, T. (2009). Simple group fMRI modeling and inference. *Neuroimage*, 47(4):1469–1475.

Nichols, T. and Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, 12(5):419–446.

Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping*, 15(1):1–25.

Nordhausen, K., Cardoso, J. F., Oja, H., and Ollila, E. (2011). *JADE: JADE and ICA performance criteria*. R package version 1.0-4.

Penny, W. D., Holmes, A., and Friston, K. (2003). *Human Brain Function*, chapter 12, pages 843–850. Academic Press London.

Penny, W. D., Trujillo-Barreto, N. J., and Friston, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24(2):350–362.

Pollard, D. (2001). Chapter 13 from Asymptopia work-in-progress.

Risk, B. B., Matteson, D. S., Ruppert, D., Eloyan, A., and Caffo, B. S. (2014). An evaluation of independent component analyses with an application to resting-state fMRI. *Biometrics*, 70(1):224–236.

Samworth, R. J. and Yuan, M. (2012). Independent component analysis via nonparametric maximum likelihood estimation. *The Annals of Statistics*, 40(6):2973–3002.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, pages 110–114.

Searle, S. R., Casella, G., and McCulloch, C. E. (2009). *Variance components*, volume 391. John Wiley & Sons.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 683–690.

Shumway, R. H. and Stoffer, D. S. (2010). *Time series analysis and its applications: with R examples*. Springer Science & Business Media.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 106(31):13040–13045.

Sporns, O. (2011). The human connectome: a complex network. *Annals of the New York Academy of Sciences*, 1224(1):109–125.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Tabelow, K., Piëch, V., Polzehl, J., and Voss, H. U. (2009). High-resolution fMRI: Overcoming the signal-to-noise problem. *Journal of neuroscience methods*, 178(2):357–365.

Tichavsky, P. and Koldovsky, Z. (2004). Optimal pairing of signal components separated by blind techniques. *Signal Processing Letters, IEEE*, 11(2):119–122.

Tichavsky, P., Koldovsky, Z., and Oja, E. (2005). Asymptotic performance of the FastICA algorithm for independent component analysis and its improvements. In *Statistical Signal Processing, 2005 IEEE/SP 13th Workshop on*, pages 1084–1089. IEEE.

Tierney, L., Rossini, A. J., Li, N., and Sevcikova, H. (2011). *snow: Simple Network of Workstations*. R package version 0.3-8.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Tohka, J., Foerde, K., Aron, A. R., Tom, S. M., Toga, A. W., and Poldrack, R. A. (2008). Automatic independent component labeling for artifact removal in fMRI. *Neuroimage*, 39(3):1227–1245.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419.

Veer, I. M., Beckmann, C. F., Van Tol, M. J., Ferrarini, L., Milles, J., Veltman, D. J., Aleman, A., Van Buchem, M. A., van der Wee, N. J., and Rombouts, S. A. (2010). Whole brain resting-state analysis reveals decreased functional connectivity in major depression. *Frontiers in systems neuroscience*, 4:1–10.

Welvaert, M., Durnez, J., Moerkerke, B., Verdoolaege, G., and Rosseel, Y. (2011). neuRosim: An R package for generating fMRI data. *Journal of Statistical Software*, 44(10):1–18.

Woolrich, M. W., Behrens, T. E., Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2004). Multilevel linear modelling for fMRI group analysis using Bayesian inference. *Neuroimage*, 21(4):1732–1747.

Worsley, K. J., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., and Evans, A. (2002). A general statistical analysis for fmri data. *Neuroimage*, 15(1):1–15.

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., Evans, A. C., et al.

(1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human brain mapping*, 4(1):58–73.

Xu, L., Johnson, T. D., Nichols, T. E., and Nee, D. E. (2009). Modeling inter-subject variability in fMRI activation location: A Bayesian hierarchical spatial model. *Biometrics*, 65(4):1041–1051.

Zhang, L., Guindani, M., and Vannucci, M. (2015). Bayesian models for functional magnetic resonance imaging data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1):21–41.